Remington's Pharmaceutical Sciences

Eighteenth Edition

Remington's Pharmaceutical Sciences

18

18TH
EDITION

Remington's

ALFONSO R GENNARO
Editor, and Chairman
of the Editorial Board

Pharmaceutical Sciences

1990

MACK PUBLISHING COMPANY
Easton, Pennsylvania 18042

Library of Congress Catalog Card No. 60-53334

ISBN 0-912734-04-3

The use of structural formulas from USAN and the USP Dictionary of Drug Names is by permission of The USP Convention. The Convention is not responsible for any inaccuracy contained herein.

NOTICE—This text is not intended to represent, nor shall it be interpreted to be, the equivalent of or a substitute for the official United States Pharmocopeia (USP) and/or the National Formulary (NF). In the event of any difference or discrepancy between the current official USP or NF standards of strength, quality, purity, packaging and labeling for drugs and representations of them herein, the context and effect of the official compendia shall prevail.

Printed in the United States of America by the Mack Printing Company, Easton, Pennsylvania

Remington's Pharmaceutical Sciences . . . *a treatise on the theory and practice of the pharmaceutical sciences, with essential information about pharmaceutical and medicinal agents; also a guide to the professional responsibilities of the pharmacist as the drug-information specialist of the health team . . . A textbook and reference work for pharmacists, physicians and other practitioners of the pharmaceutical and medical sciences.*

Eighteenth Edition—1990

Published in the 170th year of the
PHILADELPHIA COLLEGE OF PHARMACY AND SCIENCE

Remington Historical / Biographical Data

The following is a record of the editors and the dates of publication of successive editions of this book, prior to the 13th Edition known as *Remington's Practice of Pharmacy* and subsequently as *Remington's Pharmaceutical Sciences.*

First Edition, 1886
Second Edition, 1889
Third Edition, 1897
Fourth Edition, 1905

Joseph P Remington

Fifth Edition, 1907
Sixth Edition, 1917

Joseph P Remington
Assisted by
E Fullerton Cook

Seventh Edition, 1926

Editors
E Fullerton Cook
Charles H LaWall

Eighth Edition, 1936

Editors
E Fullerton Cook
Charles H LaWall

Associate Editors
Ivor Griffith
Adley B Nichols
Arthur Osol

Ninth Edition, 1948

Editors
E Fullerton Cook
Eric W Martin

Tenth Edition, 1951

Editors
E Fullerton Cook
Eric W Martin

Eleventh Edition, 1956

Editors
Eric W Martin
E Fullerton Cook

Associate Editors
E Emerson Leuallen
Arthur Osol
Linwood F Tice
Clarence T Van Meter

Twelfth Edition, 1961

Editors
Eric W Martin
E Fullerton Cook
E Emerson Leuallen
Arthur Osol
Linwood F Tice
Clarence T Van Meter

Assistant to the Editors
John E Hoover

Thirteenth Edition, 1965

Editor-in-Chief
Eric W Martin

Managing Editor
John E Hoover

Editors

Grafton D Chase	Robert E King
Herald R Cox	E Emerson Leuallen
Richard A Deno	Arthur Osol
Alfonso R Gennaro	Ewart A Swinyard
Stewart C Harvey	Clarence T Van Meter

Fourteenth Edition, 1970

Chairmon, Editorial Board
Arthur Osol

Managing Editor
John E Hoover

Editors

Grafton D Chase	Robert E King
Richard A Deno	Alfred N Martin
Alfonso R Gennaro	Ewart A Swinyard
Melvin R Gibson	Clarence T Van Meter
Stewart C Harvey	Bernard Witlin

Fifteenth Edition, 1975

Chairman, Editorial Board
Arthur Osol

Managing Editor
John E Hoover

Editors

John T Anderson	C Boyd Granberg
Cecil L Bendush	Stewart C Harvey
Grafton D Chase	Robert E King
Alfonso R Gennaro	Alfred N Martin
Melvin R Gibson	Ewart A Swinyard

Sixteenth Edition, 1980

Chairman, Editorial Board
Arthur Osol

Editors

Grafton D Chase	Robert E King
Alfonso R Gennaro	Alfred N Martin
Melvin R Gibson	Ewart A Swinyard
C Boyd Granberg	Gilbert L Zink
Stewart C Harvey	

Seventeenth Edition, 1985

Chairman, Editorial Board
Alfonso R Gennaro

Editors

Grafton D Chase	Edward G Rippie
Ara Der Marderosian	Joseph D Schwartz
Stewart Harvey	Ewart A Swinyard
Daniel A Hussar	Gilbert L Zink
Thomas Medwick	

Editorial Board Members and Editors

Alfonso R Gennaro, PhD / *Philadelphia College of Pharmacy and Science*—Professor of Chemistry. Chairman of the Editorial Board and Editor, *Remington's Pharmaceutical Sciences.* Coauthor, Chapter 22. Coeditor, Part 6, *Pharmaceutical and Medicinal Agents.*

Grofton D Chose, PhD / *Philadelphia College of Pharmacy and Science*—Emeritus Professor of Chemistry, Editor, Part 5, *Radioisotopes in Pharmacy and Medicine.* Author, Chapters 32 and 33.

Aro DerMorderosian, PhD / *Philadelphia College of Pharmacy and Science*—Professor of Pharmacognosy. Research Professor in Medicinal Chemistry. Editor, Part 1, *Orientation.* Author, Chapters 64, 74 and 96.

Stewart C Horvey, PhD / *University of Utah School of Medicine*—Professor of Pharmacology. Editorial Board Member. Editor, Part 6, *Pharmaceutical and Medicinal Agents.* Author, Chapters 35, 38, 40, 43 to 47, 49, 61 and 62. Coauthor, Chapters 36, 41 and 50.

Daniel A Hussar, PhD / *Philadelphia College of Pharmacy and Science*—Remington Professor of Pharmacy. Editorial Board Member. Editor, Part 9, *Pharmaceutical Practice.* Author, Chapters 100 and 102.

Thomas Medwick, PhD / *Rutgers University*—Professor and Chairman, Department of Pharmaceutical Chemistry. Editorial Board member. Editor, Part 3, *Pharmaceutical Chemistry,* and Part 4, *Testing and Analysis.* Coauthor, Chapter 26.

Edward G Ripple, PhD / *University of Minnesota College of Pharmacy*—Professor of Pharmaceutics. Editorial Board Member. Editor, Part 2, *Pharmaceutics.* Author, Chapter 12. Coauthor, Chapter 88.

Joseph D Schwartz, PhD / *Philadelphia College of Pharmacy and Science*—Linwood Tice Professor of Pharmaceutics. Editorial Board Member. Editor, Part 8, *Pharmaceutical Preparations and Their Manufacture.* Coauthor, Chapters 88 and 89.

Ewart A Swinyard, PhD / *University of Utah*—Professor Emeritus of Pharmacology, College of Pharmacy and School of Medicine. Editorial Board member. Editor, Part 6, *Pharmaceutical and Medicinal Agents.* Author, Chapters 39, 42, 48, 53 to 60, 63, 65 and 70. Coauthor, Chapter 66.

Gilbert L Zink, PhD / *Philadelphia College of Pharmacy and Science*—Associate Professor of Biology, Editor, Part 7, *Biological Products.* Author, Chapter 71.

Authors

The following contributors to the Eighteenth Edition of *Remington's Pharmaceutical Sciences* served as authors or coauthors, along with the editors and members of the Editorial Board, of the 109 chapters of this book.

Hamed M Abdou, PhD / Vice President, Worldwide Pharmaceutical Technical Operations, E R Squibb & Sons, Inc; Author of Chapter 30, *Instrumental Methods of Analysis* and Chapter 31, *Dissolution*.

Anu B Amerson, PharmD / Professor, College of Pharmacy/Director, Drug Information Center, Chandler Medical Center, University of Kentucky; Author of Chapter 103, *Clinical Drug Literature*.

Howard C Ansel, PhD / Professor of Pharmacy and Dean, College of Pharmacy, University of Georgia; Author of Chapter 101, *The Prescription*.

Kenneth E Avis, DSc / Emeritus Professor, Pharmaceutics, College of Pharmacy, University of Tennessee, Memphis; Author of Chapter 84, *Parenteral Preparations*.

Leonard C Bailey, PhD / Associate Professor of Pharmaceutical Chemistry, Rutgers University College of Pharmacy; Author of Chapter 29, *Chromatography*.

Lawrence H Block, PhD / Professor of Pharmaceutics, Duquesne University School of Pharmacy; Author of Chapter 87, *Medicated Applications*.

Joseph B Bogardus, PhD / Basic Pharmaceutics Research, Bristol-Myers Company; Coauthor of Chapter 18, *Reaction Kinetics*.

Sanford Bolton, PhD / Chairman, Department of Pharmacy and Administrative Sciences, St John's University; Author of Chapter 10, *Statistics*.

John Bosso, PharmD / Professor of Clinical Pharmacy and Adjunct Professor of Pediatrics, College of Pharmacy and School of Medicine, University of Utah; Coauthor of Chapter 34, *Diseases: Manifestations and Pathophysiology*.

B Sue Brizuela, MS / Assistant Professor of Information Science, Head of Public Services, Joseph W England Library, Philadelphia College of Pharmacy and Science; Coauthor of Chapter 7, *Drug Information*.

Dale B Christensen, PhD / Associate Professor, Department of Pharmacy Practice, School of Pharmacy, University of Washington; Coauthor of Chapter 11, *Computer Science*.

Sebastian G Ciancio, DDS / Professor and Chairman, Department of Periodontology, School of Dental Medicine, State University of New York at Buffalo; Author of Chapter 109, *Dental Services*.

Kenneth A Connors, PhD / Professor of Pharmaceutics, School of Pharmacy, University of Wisconsin; Author of Chapter 14, *Complex Formation*.

Anthony J Cutie, PhD / Professor of Pharmaceutics, Arnold and Marie Schwartz College of Pharmacy and Health Sciences, Long Island University; Coauthor of Chapter 92, *Aerosols*.

Anthony R DiSanto, PhD / Vice President, Drug Delivery Research and Development, The Upjohn Company; Author of Chapter 76, *Bioavailability and Bioequivalency Testing*.

Clarence A Discher,* PhD / Professor Emeritus, Rutgers University; Author of Chapter 21, *Inorganic Pharmaceutical Chemistry*.

Clyde R Erskine, Jr, BSc / Vice President, Corporate Quality Audits and Services, SmithKline Beckman Corporation; Author of Chapter 82, *Quality Assurance and Control*.

Lorraine D Evans, BS, H(ASCP) / Clinical Pathology, Bristol-Myers Company; Coauthor of Chapter 28, *Clinical Analysis*.

William E Fassett, BS, MBA / Assistant Professor, Department of Pharmacy Practice, School of Pharmacy, University of Washington; Coauthor of Chapter 11, *Computer Science*.

Joseph L Fink III, BS(Pharm), JD / Assistant Dean and Professor, College of Pharmacy, University of Kentucky; Coauthor of Chapter 107, *Laws Governing Pharmacy*.

Michael R Franklin, PhD / Professor of Pharmacology, College of Pharmacy and School of Medicine, University of Utah; Author of Chapter 52, *Enzymes*.

Ruta Freimanis, BS, RPh / Associate Secretary, United States Adopted Names Council; Coauthor of Chapter 24, *Drug Nomenclature—United States Adopted Names*.

James W Freston, MD, PhD / Professor and Chairman, Department of Medicine, University of Connecticut Health Center; Coauthor of Chapter 34, *Diseases: Manifestations and Pathophysiology*.

Robert L Giles, BA / Vice President and General Manager, Glenn Beall Engineering Inc; Coauthor of Chapter 80, *Plastic Packaging Materials*.

Harold N Godwin, MS / Professor and Director of Pharmacy, The University of Kansas Medical Center; Author of Chapter 94, *Institutional Patient Care*.

Frederick J Goldstein, PhD / Professor of Pharmacology, Philadelphia College of Pharmacy and Science; Coauthor of Chapter 69, *Pharmacological Aspects of Substance Abuse*.

A Richard Goolkasian, BS, RPh / Director of Alumni and Professional Affairs, Massachusetts College of Pharmacy and Allied Health Sciences; Author of Chapter 1, *Scope*.

Gerald Hecht, PhD / Director Process Development, Alcon Laboratories; Coauthor of Chapter 86, *Ophthalmic Preparations*.

Judith A Hesp, MS / Instructor in Information Science, Coordinator of Bibliographic Instruction, Joseph W England Library, Philadelphia College of Pharmacy and Science; Coauthor of Chapter 7, *Drug Information*.

Gregory J Higby, PhD / Director, American Institute of the History of Pharmacy, School of Pharmacy, University of Wisconsin-Madison; Author of Chapter 2, *Evolution of Pharmacy*.

Andrew S Katocks, Jr, PhD / Senior Research Pharmacologist, American Cyanamid Company, Medical Research Division; Coauthor of Chapter 27, *Biological Testing*.

Calvin H Knowlton, MDiv, RPh / Clinical Associate Professor of Pharmacy, Philadelphia College of Pharmacy and Science; Author of Chapter 4, *The Practice of Community Pharmacy*.

Richard W Knueppel, RPh / President, Knueppel Home Health Care Center; Author of Chapter 104, *Health Accessories*.

Harry B Kostenbauder, PhD / Associate Dean for Research, College of Pharmacy, University of Kentucky; Coauthor of Chapter 18, *Reaction Kinetics*.

Richard L Kronenthal, PhD / Director of Research, Ethicon Inc; Author of Chapter 105, *Surgical Supplies*.

Arthur J Lawrence, PhD, RPh / Office of the Assistant Secretary of Health, US Public Health Service; Author of Chapter 6, *Pharmacists in Government*.

Eric J Lien, PhD / Professor of Pharmacy / Pharmaceutics and Biomedical Chemistry, School of Pharmacy, University of Southern California; Author of Chapter 13, *Molecular Structure, Properties and States of Matter*.

Mark A Longer, PhD / MCR Research Fellow, Department of Biological Sciences, University of Keele; Coauthor of Chapter 91, *Sustained-Release Drug Delivery Systems*.

Werner Lowenthal, PhD / Professor of Pharmacy and Pharmaceutics and Professor of Educational Development and Planning, School of Pharmacy, Medical College of Virginia; Author of Chapter 9, *Metrology and Calculation*.

Karen B Main, PhD / Physical Pharmacist, Pharmaceutical Development Department, ICI Pharmaceuticals Group; Coauthor of Chapter 26, *Analysis of Medicinals*.

Duane D Miller, PhD / Professor and Chairman, Division of Medicinal Chemistry and Pharmacognosy, College of Pharmacy,

* Deceased

The Ohio State University; Author of Chapter 25, *Structure-Activity Relationship and Drug Design.*

Michael Montagne, PhD / Associate Professor of Pharmacy Administration, Philadelphia College of Pharmacy and Science; Coauthor of Chapter 3, *Ethics* and Author of Chapter 99, *Drug Education.*

John D Mullins, PhD / Consultant; Coauthor of Chapter 86, *Ophthalmic Preparations.*

Maven J Myers, PhD / Professor of Pharmacy Administration, Philadelphia College of Pharmacy and Science; Coauthor of Chapter 3, *Ethics.*

J G Nairn, PhD / Professor of Pharmacy, Faculty of Pharmacy, University of Toronto; Author of Chapter 83, *Solutions, Emulsions, Suspensions and Extracts.*

Paul J Niebergall, PhD / Professor of Pharmaceutical and Sciences / Director, Pharmaceutical Development Center, Medical University of South Carolina; Author of Chapter 17, *Ionic Solutions and Electrolytic Equilibria.*

Robert E O'Connor, PhD / Merck Frosst Canada, Inc; Coauthor of Chapter 88, *Powders.*

Melanie O'Neill / Becton Dickinson & Company; Coauthor of Chapter 78, *Sterilization.*

Richard W Pecina, PhD / President, Richard W Pecina & Associates; Coauthor of Chapter 80, *Plastic Packaging Materials.*

Garnet E Peck, PhD / Professor of Industrial Pharmacy / Director of the Industrial Pharmacy Laboratory, Purdue University; Author of Chapter 77, *Separation.*

G Briggs Phillips, PhD / Becton Dickinson & Company; Coauthor of Chapter 78, *Sterilization.*

Nicholas G Popovich, PhD / Associate Professor of Pharmacy Practice, School of Pharmacy and Pharmacal Sciences, Purdue University; Author of Chapter 93, *Ambulatory Patient Care.*

Stuart C Porter, PhD / Vice President, Research and Development, Colorcon; Author of Chapter 90, *Coating of Pharmaceutical Dosage Forms.*

Galen Radebaugh, PhD / Director of Pharmaceutics, Parke-Davis Pharmaceutical Research Division, Warner-Lambert Company; Coauthor of Chapter 75, *Preformulation.*

Paul L Ranelli, PhD / Assistant Professor of Pharmacy Administration, Philadelphia College of Pharmacy and Science; Author of Chapter 98, *Patient Communication.*

Louis J Ravin, PhD / Department of Pharmaceutics, Research and Development, Smith Kline & French Laboratories; Coauthor of Chapter 75, *Preformulation.*

Jack W Reich, PhD / Vice President Regulatory Affairs, Gensia Pharmaceuticals, Inc; Author of Chapter 8, *Research.*

James W Richards, MBA / Professor of Pharmacy Administration, College of Pharmacy, University of Michigan; Author of Chapter 108, *Community Pharmacy Economics and Management.*

Jack Robbins, PhD / Director, Pharmacy Affairs, Schering Laboratories; Author of Chapter 5, *Opportunities for Pharmacists in the Pharmaceutical Industry.*

Joseph R Robinson, PhD / Professor of Pharmacy, School of Pharmacy, University of Wisconsin; Coauthor of Chapter 91, *Sustained-Release Drug Delivery Systems.*

Frank Roia, PhD / Professor of Biology, Philadelphia College of Pharmacy and Science; Author of Chapter 72, *Immunizing Agents and Diagnostic Skin Antigens.*

Douglas E Rollins, MD, PhD / Associate Professor of Medicine and Pharmacology, School of Medicine and College of Pharmacy, University of Utah; Author of Chapter 37, *Clinical Pharmacokinetics.*

G Victor Rossi, PhD / Vice Presidont of Academic Affairs / Professor of Pharmacology, Philadelphia College of Pharmacy and Science; Coauthor of Chapter 27, *Biological Testing* and Coauthor of Chapter 69,

Edward Rudnic, PhD / Director, Formulation Development, Schering Research; Coauthor of Chapter 89, *Oral Solid Dosage Forms.*

Donald O Schiffman, PhD / Secretary, United States Adopted Names Council; Coauthor of Chapter 24, *Drug Nomenclature—United States Adopted Names.*

Hans Schott, PhD / Professor of Pharmaceutics and Colloid Chemistry, School of Pharmacy, Temple University; Coauthor of Chapter 19, *Disperse Systems* and Author of Chapter 20, *Rheology.*

John J Sciarra, PhD / President, Retail Drug Institute / Professor of Industrial Pharmacy, Arnold and Marie Schwartz College of Pharmacy and Health Sciences, Long Island University; Coauthor of Chapter 92, *Aerosols.*

John H Shinkai, PhD / Emeritus Professor of Pharmaceutical Chemistry, Rutgers University, College of Pharmacy; Coauthor of Chapter 22, *Organic Pharmaceutical Chemistry.*

E Richard Shough, PhD / Associate Dean and Professor, University of Oklahoma Health Sciences Center, College of Pharmacy; Author of Chapter 73, *Allergenic Extracts.*

Frederick P Siegel, PhD / Professor of Pharmaceutics, College of Pharmacy, University of Illinois; Author of Chapter 79, *Tonicity, Osmoticity, Osmolality and Osmolarity.*

Larry M Simonsmeier, BS(Pharm), JD / Associate Dean and Professor, College of Pharmacy, Washington State University; Coauthor of Chapter 107, *Laws Governing Pharmacy.*

Robert D Smyth, PhD / Vice President, Pharmaceutical Development, Bristol-Myers Company; Coauthor of Chapter 28, *Clinical Analysis.*

Thomas C Snader, PharmD / Consultant Pharmacist; Author of Chapter 95, *Long-Term Care Facilities.*

Theodore D Sokoloski, PhD / Professor of Pharmacy, College of Pharmacy, The Ohio State University; Author of Chapter 16, *Solutions and Phase Equilibria.*

Robert B Stewart, MS / Professor and Chairman, Department of Pharmacy Practice, College of Pharmacy, University of Florida; Author of Chapter 67, *Adverse Drug Reactions.*

James Swarbrick, DSc, PhD / Professor and Chairman, Division of Pharmaceutics, School of Pharmacy, University of North Carolina at Chapel Hill; Coauthor of Chapter 19, *Disperse Systems.*

Anthony R Temple, MD / Director, Regulatory and Medical Affairs, McNeil Consumer Products Company / Adjunct Associate Professor, Department of Pediatrics, University of Pennsylvania School of Medicine / Lecturer, Philadelphia College of Pharmacy and Science; Author of Chapter 106, *Paison Control.*

John P Tischio, PhD / Principle Scientist, Immunohiology Research Institute; Author of Chapter 68, *Pharmacogenetics.*

Salvatore J Turco, PharmD / Professor of Pharmacy, Temple University School of Pharmacy; Author of Chapter 85, *Intravenous Admixtures.*

Elizabeth B Vadas, PhD / Merck Frosst Canada, Inc; Author of Chapter 81, *Stability of Pharmaceutical Products.*

Ernestine Vanderveen, PhD / National Institute on Drug Abuse, ADAMHA; Coauthor of Chapter 51, *Vitamins and Other Nutrients.*

John E Vandervcen, PhD / Division of Nutrition, Food and Drug Administration; Coauthor of Chapter 51, *Vitamins and Other Nutrients.*

Vincent S Venturella, PhD / Section Manager, Pharmaceutical Research, Anaquest, Div of BOC; Author of Chapter 23, *Natural Products.*

Albert I Wertheimer, PhD / Professor and Director, Department of Graduate Studies in Social and Administrative Pharmacy, College of Pharmacy, University of Minnesota; Author of Chapter 97, *The Patient: Behavioral Determinants.*

Timothy S Wiedmann, PhD / Assistant Professor, College of Pharmacy, University of Minnesota; Author of Chapter 15, *Thermodynamics.*

C Dean Withrow, PhD / Associate Professor of Pharmacology, School of Medicine, University of Utah; Coauthor of Chapter 36, *Basic Pharmacokinetics*, Coauthor of Chapter 41, *Cardiovascular Drugs* and Coauthor of Chapter 50, *Hormones.*

George Zografi, PhD / Professor, School of Pharmacy, University of Wisconsin; Coauthor of Chapter 19, *Disperse Systems.*

Preface to the First Edition

The rapid and substantial progress made in Pharmacy within the last decade has created a necessity for a work treating of the improved apparatus, the revised processes, and the recently introduced preparations of the age.

The vast advances made in theoretical and applied chemistry and physics have much to do with the development of pharmaceutical science, and these have been reflected in all the revised editions of the Pharmacopoeias which have been recently published. When the author was elected in 1874 to the chair of Theory and Practice of Pharmacy in the Philadelphia College of Pharmacy, the outlines of study which had been so carefully prepared for the classes by his eminent predecessors, Professor William Procter, Jr, and Professor Edward Parrish, were found to be not strictly in accord, either in their arrangement of the subjects or in their method of treatment. Desiring to preserve the distinctive characteristics of each, an effort was at once made to frame a system which should embody their valuable features, embrace new subjects, and still retain that harmony of plan and proper sequence which are absolutely essential to the success of any system.

The strictly alphabetical classification of subjects which is now universally adopted by pharmacopoeias and dispensatories, although admirable in works of reference, presents an effectual stumbling block to the acquisition of pharmaceutical knowledge through systematic study; the vast accumulation of facts collected under each head being arranged lexically, they necessarily have no connection with one another, and thus the saving of labor effected by considering similar groups together, and the value of the association of kindred subjects, are lost to the student. In the method of grouping the subjects which is herein adopted, the constant aim has been to arrange the latter in such a manner that the reader shall be gradually led from the consideration of elementary subjects to those which involve more advanced knowledge, whilst the groups themselves are so placed as to follow one another in a natural sequence.

The work is divided into six parts. Part I is devoted to detailed descriptions of apparatus and definitions and comments on general pharmaceutical processes.

The Official Preparations alone are considered in Part II. Due weight and prominence are thus given to the Pharmacopoeia, the National authority, which is now so thoroughly recognized.

In order to suit the convenience of pharmacists who prefer to *weigh solids* and *measure liquids*, the official formulas are expressed, in addition to parts by weight, in *avoirdupois weight* and *apothecaries' measure*. These equivalents are printed in *bold type* near the margin, and arranged so as to fit them for quick and accurate reference.

Part III treats of Inorganic Chemical Substances. Precedence is of course given to official preparation in these. The descriptions, solubilities, and tests for identity and impurities of each substance are systematically tabulated under its proper title. It is confidently believed that by this method of arrangement the valuable descriptive features of the Pharmacopoeia will be more prominently developed, ready reference facilitated, and close study of the details rendered easy. Each chemical operation is accompanied by equations, whilst the reaction is, in addition, explained in words.

The Carbon Compounds, or Organic Chemical Substances, are considered in Part IV. These are naturally grouped according to the physical and medical properties of their principal constituents, beginning with simple bodies like cellulin, gum, etc, and progressing to the most highly organized alkaloids, etc.

Part V is devoted to Extemporaneous Pharmacy. Care has been taken to treat of the practice which would be best adapted for the needs of the many pharmacists who conduct operations upon a moderate scale, rather than for those of the few who manage very large establishments. In this, as well as in other parts of the work, operations are illustrated which are conducted by manufacturing pharmacists.

Part VI contains a formulary of Pharmaceutical Preparations which have not been recognized by the Pharmacopoeia. The recipes selected are chiefly those which have been heretofore rather difficult of access to most pharmacists, yet such as are likely to be in request. Many private formulas are embraced in the collection; and such of the preparations of the old Pharmacopoeias as have not been included in the new edition, but are still in use, have been inserted.

In conclusion, the author ventures to express the hope that the work will prove an efficient help to the pharmaceutical student as well as to the pharmacist and the physician. Although the labor has been mainly performed amidst the harassing cares of active professional duties, and perfection is known to be unattainable, no pains have been spared to discover and correct errors and omissions in the text. The author's warmest acknowledgments, are tendered to Mr A B Taylor, Mr Joseph McCreery, and Mr George M Smith for their valuable assistance in revising the proof sheets, and to the latter especially for his work on the index. The outline illustrations, by Mr John Collins, were drawn either from the actual objects or from photographs taken by the author.

Philadelphia, October, 1885 JPR.

Preface to the Eighteenth Edition

In anticipation of setting forth this *Preface* and prior to gathering thoughts on paper (or more accurately, the word processor!), this Editor paused to reread the preface to the first edition of *Remington*, published in 1885. Since it appears on the preceding page of this book it is recommended highly. The first paragraph would he just as suitable today as penned by Professor Remington 105 years ago.

Each decade transcends the previous one and the pharmaceutical and health sciences are not laggards. Every revision of *Remington* has encompassed new viewpoints, ideas, doctrines or principles which, perhaps, were inconceivable for the previous edition. It is a tribute to the authors and editors that they have kept abreast of the burgeoning literature in their respective fields of expertise.

Change not withstanding, the organization of this edition is similar to its immediate predecessors, being divided into 9 Parts, each subdivided into several chapters. Every chapter has been culled, revised and rewritten to update the material presented.

Two new chapters are evident; *Biotechnology and Drugs* (Chapter 74) and *Drug Education* (Chapter 99). Three chapters of the previous edition, which embraced *Interfacial and Particle Phenomena* and *Colloidal and Coarse Dispersions* have been winnowed and combined into a single chapter entitled, *Disperse Systems* (Chapter 19).

The current revision contains an additional 21 pages. A large amount of space (about 19 pages) gleaned from the review and condensation process, coupled with the extra pages, have been devoted primarily to expanding the contents of Part 6, *Pharmaceutical and Medicinal Agents* and Part 9, *Pharmaceutical Practice*.

Excessive duplication of text is the bane of any editor dealing with a multitude of authors. While some duplication in the discussion of rudimentary concepts is beneficial, there has been a special effort to cross-reference and eliminate unnecessary repetition. Space is at such a premium that it is hoped the reader will not be offended by being diverted to a different section of the text in order to obtain supplementary information.

Photographs which depicted the typical "black box" have been eliminated almost completely and replaced by line drawings or schematic diagrams which are instructive rather than picturesque.

Most of the drug monographs have been revamped not only as a means of updating, but to gain a degree of uniformity. All structural formulas are now in the standard *USAN* form. Duplication of chemical names has been minimized and the inclusion of trade names increased. No attempt has been made to ferret out every trade name by which a product is known, and only the most common are mentioned. The standard format for the major monographs is: Official Name, chemical name (*CAS*—inverted), trade name(s) and manufacturer(s), structural formula, *CAS* (*Chemical Abstracts System*) registry numher (in brackets), molecular formula and formula weight (in parenthesis). This is followed by the method of preparation (or a reference if the method is lengthy), physical description, solubility, uses, dose and dosage forms.

The number of authors remains at 97, however, 36 new authors have joined as contributors to *Remington*. As the credentials of the new authors touch upon many areas of pharmacy, every section of the book has been invigorated by the incorporation of updated and fresh concepts.

As one primarily responsible for the production of a comprehensive text devoted to the science and practice of pharmacy, the wisdom of Dr Eric Martin, editor of the 13th Edition, in creating an Editorial Board to share the enormous burden, has been evident constantly. Each of the section editors labored diligently to comply with the logistics of maintaining a smooth flow of manuscripts and proofs. Also, each section editor doubled as an author or coauthor of one or more chapters. It would be remiss not to extend special mention to this group of dedicated people.

Four members of the Editorial Board are serving for the first time after having been authors for several editions. Dr Ara DerMarderosian of PCP&S, Editor for Part 1; Dr Daniel Hussar, also of PCP&S, Part 9; Dr Edward Rippie of the University of Minnesota, Part 2; and Dr Joseph Schwartz of PCP&S, Part 8. Each of the new members literally "jumped into the fray," gave much of their precious time and have become "blooded" members of the staff.

The stalwarts of the Editorial Board surviving the tribulations of one or more previous editions of this work demand singular attention. Dr Grafton Chase of PCP&S for Part 5, *Radioisotopes in Pharmacy & Medicine*; Dr Thomas Medwick of Rutgers University for Part 3, *Pharmaceutical Chemistry* and Part 4, *Testing and Analysis*; and Dr Gilbert Zink of PCP&S for Part 7, *Biological Products*.

Two dauntless, prolific contributors claim special recognition. Drs Stewart Harvey and Ewart Swinyard, both of the University of Utah, have served on the Editorial Board for twenty and twenty-five years respectively. They bear the burden of Part 6, *Pharmaceutical & Medicinal Agents*, which comprises over one-third of the book. Their diligence and meticulous attention to detail has eased the task of this Editor. Our relationship over the past several decades has been one of exceptional pleasure.

The Mack Publishing Company, through Messers Paul Mack and David Palmer, continues its unrelenting support, which has endured through many, many editions of this publication. Special commendation must be extended to Ms Nancy Smolock, of the Mack organization, as she was the person who interfaced with the Editorial Board. She was competent, cooperative and much too tolerant of the many requests made of her.

As with any publication a few of the editorial staff bear the hrunt of the unglamorous, but absolutely essential, chores associated with the production of this voluminous tome. It mandates a close working relationship and, at times, restraint and concession to sustain the harmony necessary to function efficiently. One often encounters the aphorism usually attributed to administrators, "When three managers meet to discuss a problem there arise four points-of-view." Fortunately, this dilemma did not surface in the association of this Editor with Mr John Hoover and Ms Bonnie Packer.

After shepherding this publication through four editions, the Twelfth to Fifteenth, following a short hiatus for the Sixteenth, Mr Hoover returned in a lesser capacity with the Seventeenth revision. With the current edition he reassumes the role of Managing Editor and his experience in pharmacy, journalism and the publishing business, have provided the capabilities needed to translate a disarranged manuscript into a format acceptable hy the publisher and pleasing to the reader.

Ms Packer accepted the assignment of scrutinizing every word of text in the proof stages. Combining her skills in the

health and social sciences, she assumed the charge of reading primarily for comprehension and clarity of presentation, while concurrently uncovering typographical, spelling and grammatical errors which, although unpardonable, are ever-present. As a consequence of her deliberations, passages were often rephrased and refined to portray a concept from the viewpoint of the student, for whom this work primarily is directed.

The *Index* was developed by Mr Hoover, with the assistance of Ms Packer. Much use was made of the computer in ensuring that a complete, practical and useful index was created. It is the opinion of this Editor that a major weakness encountered in most reference books is a perfunctory, casual index which amounts to little more than an expanded table of contents. Users of the index of this book will find it "friendly."

Philadelphia, February, 1990 ARG

Table of Contents

CHAPTER 16

Solutions and Phase Equilibria

Theodore D Sokoloski, PhD
Professor of Pharmacy, College of Pharmacy
Ohio State University
Columbus, OH 43210

Solutions and Solubility

A solution is a chemically and physically homogeneous mixture of two or more substances. The term solution generally denotes a homogeneous mixture that is liquid even though it is possible to have homogeneous mixtures which are solid or gaseous. Thus, it is possible to have solutions of solids in liquids, liquids in liquids, gases in liquids, gases in gases and solids in solids. The first three of these are most important in pharmacy and ensuing discussions will be concerned primarily with them.

In pharmacy different kinds of liquid dosage forms are used and all consist of the dispersion of some substance or substances in a liquid phase. Depending on the size of the dispersed particle they are classified as *true solutions*, *colloidal solutions* or *suspensions*. If sugar is dissolved in water, it is supposed that the ultimate sugar particle is of molecular dimensions and that a *true solution* is formed. On the other hand, if very fine sand is mixed with water, a *suspension* of comparatively large particles, each consisting of many molecules, is obtained. Between these two extremes lie *colloidal solutions*, the dispersed particles of which are larger than those of true solutions but smaller than the particles present in suspensions. In this chapter only true solutions will be discussed.

It is possible to classify broadly all solutions as one of two types.

In the first type, although there may be a lesser or greater interaction between the dispersed substance (the solute) and the dispersing medium (the solvent), the solution phase contains the same chemical entity as found in the solid phase and, thus, upon removal of the solvent, the solute is recovered unchanged. One example would be sugar dissolved in water where, in the presence of sugar in excess of its solubility, there is an equilibrium between sugar molecules in the solid phase with sugar molecules in the solution phase. A second example would be dissolving silver chloride in water. Admittedly, the solubility of this salt in water is low, but it is finite. In this case the solvent contains silver and chloride ions and the solid phase contains the same material. The removal of the solvent yields initial solute.

In the second type the solvent contains a compound which is different from that in the solid phase. The difference between the compound in the solid phase and solution is due generally to some chemical reaction that has occurred in the solvent. An example would be dissolving aspirin in an aqueous solvent containing some basic material capable of reacting with the acid aspirin. Now the species in solution would not only be undissociated aspirin, but aspirin also as its anion, whereas the species in the solid phase is aspirin in only its undissociated acid form. In this situation, if the solvent were removed, part of the substance obtained (the salt of aspirin) would be different from what was present initially in the solid.

Solutions of Solids in Liquids

Reversible Solubility without Chemical Reaction—From a pharmaceutical standpoint solutions of solids in liquids, with or without accompanying chemical reaction in the solvent, are of the greatest importance, and many quantitative data on the behavior and properties of such solutions are available. This discussion will be concerned with definitions of solubility, the rate at which substances go into solution and with temperature and other factors which control solubility.

Solubility—When an excess of a solid is brought into contact with a liquid, molecules of the former are removed from its surface until equilibrium is established between the molecules leaving the solid and those returning to it. The resulting solution is said to be saturated at the temperature of the experiment, and the extent to which the solute dissolves is referred to as its *solubility*. The extent of solubility of different substances varies from almost imperceptible amounts to relatively large quantities, but for any given solute the solubility has a constant value at constant temperature.

Under certain conditions it is possible to prepare a solution containing a larger amount of solute than is necessary to form a saturated solution. This may occur when a solution is saturated at one temperature, the excess of solid solute removed and the solution cooled. The solute present in solution, even though it may be less-soluble at the lower temperature, does not always separate from the solution and there is produced a *supersaturated solution*. Such solutions, formed by sodium thiosulfate or potassium acetate, for example, may be made to deposit their excess of solute by vigorous shaking, scratching the side of the vessel in contact with the solution or introducing into the solution a small crystal of the solute.

Methods of Expressing Solubility—When quantitative data are available, solubilities may be expressed in many ways. For example, the solubility of sodium chloride in water at 25° may be stated as

1. 1 g of sodium chloride dissolves in 2.786 mL of water. (An approximation of this method is used by the USP.)
2. 35.89 g of sodium chloride dissolves in 100 mL of water.
3. 100 mL of a saturated solution of sodium chloride in water contains 31.71 g of solute.
4. 100 g of a saturated solution of sodium chloride in water contains 26.47 g of solute.
5. 1 L of a saturated solution of sodium chloride in water contains 5.425 moles of solute. This also may be stated as a saturated solution of sodium chloride in water is 5.425 molar with respect to the solute.

In order to calculate *3* from *1* or *2* it is necessary to know the density of the solution, in this case 1.198 g/mL. To calculate *5*, the number of grams of solute in 1000 mL of solution (obtained by multiplying the data in (*3*) by ten) is divided by the molecular weight of sodium chloride, namely, 58.45.

Several other concentration expressions are used. Molality is the number of moles of solute in 1000 g of solvent and could be calculated from the data in *4* by subtracting grams

The author acknowledges the kind assistance of Dr Gordon L Flynn, University of Michigan, in the revision of parts of this chapter.

Table I—Descriptive Terms for Solubility

Descriptive Terms	Parts of Solvent for 1 Part of Solute
Very soluble	Less than 1
Freely soluble	From 1 to 10
Soluble	From 10 to 30
Sparingly soluble	From 30 to 100
Slightly soluble	From 100 to 1000
Very slightly soluble	From 1000 to 10,000
Practically insoluble, or insoluble	More than 10,000

of solute from grams of solution to obtain grams of solvent, relating this to 1000 g of solvent and dividing by molecular weight to obtain moles.

Mole fraction is the fraction of the total number of moles present which are moles of one component. Mole % may be obtained by multiplying mole fraction by 100. Normality refers to the number of gram equivalent weights of solute dissolved in 1000 mL of solution.

In pharmacy, use also is made of three other concentration expressions. Percent by weight (% *w/w*) is the number of grams of solute per 100 g of solution and is exemplified by *4* above. Percent weight in volume (% *w/v*) is the number of grams of solute per 100 mL of solution and is exemplified by *3* above. Percent by volume (% *v/v*) is the number of milliliters of solute in 100 mL of solution, referring to solutions of liquids in liquids. The USP indicates that the term "percent," when unqualified, means percent weight in volume for solutions of solids in liquids and percent by volume for solutions of liquids in liquids.

When, in pharmacopeial texts, it has not been possible, or in some instances not desirable, to indicate exact solubility, a descriptive term has been used. Table I indicates the meaning of such terms.

Rate of Solution—It is possible to define quantitatively the rate at which a solute goes into solution. The simplest treatment is based on a model depicted in Fig 16-1. A solid particle dispersed in a solvent is surrounded by a thin layer of solvent having a finite thickness, l in cm. The layer is an integral part of the solid and, thus, is referred to characteristically as the "stagnant layer." This means that regardless of how fast the bulk solution is stirred the stagnant layer remains a part of the surface of the solid, moving wherever the particle moves. The thickness of this layer may get smaller as the stirring of the bulk solution increases, but it is important to recognize that this layer will always have a finite thickness however small it may get.

Using Fick's First Law of Diffusion the rate of solution of the solid can be explained, in the simplest case, as the rate at which a dissolved solute particle diffuses through the stagnant layer to the bulk solution. The driving force behind the movement of the solute molecule through the stagnant layer is the difference in concentration that exists between the concentration of the solute, C_1, in the stagnant layer at the surface of the solid and its concentration, C_2, on the farthest side of the stagnant layer (see *Diffusion in Liquids*, page 221). The greater this difference in concentration ($C_1 - C_2$), the faster the rate of solution.

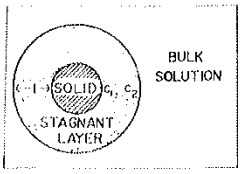


Fig 16-1. Physical model representing the dissolution process.

According to Fick's Law, the rate of solution also is directly proportional to the area of the solid, A in cm^2, exposed to solvent and inversely proportional to the length of the path through which the dissolved solute molecule must diffuse. Mathematically, then, the rate of solution of the solid is given by

$$\text{Rate of solution} = \frac{DA}{l}(C_1 - C_2) \qquad (1)$$

where D is a proportionality constant called the diffusion coefficient in cm^2/sec. In measuring the rate of solution experimentally, the concentration C_2 is maintained at a low value compared to C_1 and hence considered to have a negligible effect on the rate. Furthermore, C_1 most often is the saturation solubility of the solute. Hence Eq 1 is simplified to

$$\text{Rate of solution} = \frac{DA}{l}(\text{saturation solubility}) \qquad (2)$$

Equation 2 quantitatively explains many of the phenomena commonly observed that affect the rate at which materials dissolve.

1. Small particles go into solution faster than large particles. For a given mass of solute, as the particle size becomes smaller, the surface area per unit mass of solid increases; Eq 2 shows that as area increases, the rate must increase proportionately. Hence, if a pharmacist wishes to increase the rate of solution of a drug, its particle size should be decreased.
2. Stirring a solution increases the rate at which a solid dissolves. This is because the thickness of the stagnant layer depends on how fast the bulk solution is stirred; as stirring rate increases, the length of the diffusional path decreases. Since the rate of solution is proportional inversely to the length of the diffusional path, the faster the solution is stirred, the faster the solute will go into solution.
3. The more soluble the solute, the faster is its rate of solution. Again, Eq 2 predicts that the larger the saturation solubility, the faster the rate.
4. With a viscous liquid the rate of solution is decreased. This is because the diffusion coefficient is proportional inversely to the viscosity of the medium; the more viscous the solvent, the slower the rate of solution.

Heat of Solution and Temperature Dependency—Turning from the kinetic aspects of dissolution, this discussion will be concerned with the situation where there is thermodynamic equilibrium between solute in its solid phase and the solute in solution. (It is assumed that there is an amount of solid material in excess of the amount that can go into solution; hence, a solid phase is always present.) As defined earlier, the concentration of solute in solution at equilibrium is the saturation solubility of the substance.

When a solid (Solute A) dissolves in some solvent two steps may be considered as occurring: the solid absorbs energy to become a liquid and then the liquid dissolves.

$$A_{(\text{solid})} \rightleftarrows A_{(\text{liquid})} \rightleftarrows A_{(\text{solution})}$$

For the overall dissolution the equilibrium existing between solute molecules in the solid and solute molecules in solution may be treated as any equilibrium. Thus, for Solute A in equilibrium with its solution

$$A_{(\text{solid})} \rightleftarrows A_{(\text{solution})}$$

Using the Law of Mass Action an equilibrium constant for this system can be defined, just as any equilibrium constant may be written as

$$K_{eq} = \frac{a_{(\text{solution})}}{a_{(\text{solid})}}$$

where a denotes the activity of the solute in each phase. Since the activity of a solid is defined as unity

$$K_{eq} = a_{(\text{solution})}$$

Because the activity of a compound in dilute solution is approximated by its concentration and, because this concentration is the saturation solubility, K_S, the van't Hoff Equation (for a more complete treatment, see Ref 1, page 113) may be used, which defines the relationship between an equilibrium constant (here, solubility) and absolute temperature.

$$\frac{d \log K_S}{dT} = \frac{\Delta H}{2.3RT^2} \tag{3}$$

where $d \log K_S/dT$ is the change of $\log K_S$ with a unit change of absolute temperature, T; ΔH is a constant which in this situation is the heat of solution for the overall process (solid $\rightleftarrows$ liquid $\rightleftarrows$ solution); and R is the gas constant, 1.99 cal/mole/deg. Equation 3, a differential, may be solved to give

$$\log K_S = -\frac{\Delta H}{2.3RT} + J \tag{4}$$

where J is a constant. A more useful form of this equation is

$$\log \frac{K_{S,T_2}}{K_{S,T_1}} = \frac{\Delta H(T_2 - T_1)}{2.3RT_1T_2} \tag{5}$$

where K_{S,T_1} is the saturation solubility at absolute temperature T_1 and K_{S,T_2} is the solubility at temperature T_2. Through the use of Eq 5, if ΔH and the solubility at one temperature are known, the solubility at any other temperature can be calculated.

Effect of Temperature—As is evident from Eq 4, the solubility of a solid in a liquid depends on the temperature. If, in the process of solution, heat is absorbed (as evidenced by a reduction in temperature), ΔH is by convention positive and the solubility of the solute will increase with increasing temperature. Such is the case for most salts, as is shown in Fig 16-2, in which the solubility of the solute is plotted as the ordinate and the temperature as the abscissa, and the line joining the experimental points represents the solubility curve for that solute.

If a solute gives off heat during the process of solution (as evidenced by an increase in temperature), ΔH is, by convention, negative and solubility decreases with an increase in temperature. This is the case with calcium hydroxide and, at higher temperatures, with calcium sulfate. (Because of the slight solubility of these substances their solubility curves are not included.) When heat is neither absorbed nor given off, the solubility is not affected by variation of temperature as is nearly the case with sodium chloride.

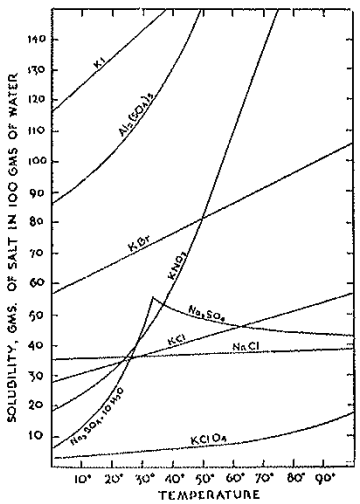


Fig 16-2. Effect of heat on solubility.

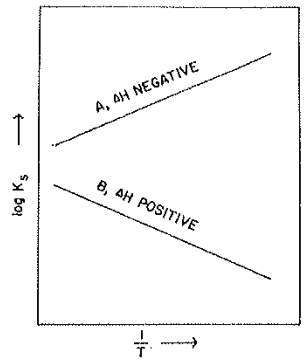


Fig 16-3. Typified relationship between the logarithm of the saturation solubility and the reciprocal of the absolute temperature.

Solubility curves usually are continuous as long as the chemical composition of the solid phase in contact with the solution remains unchanged, but if there is a transition of the solid phase from one form to another, a break will be found in the curve. Such is the case with $Na_2SO_4 \cdot 10H_2O$, which dissolves with absorption of heat up to a temperature of 32.4°, at which point there is a transition of the solid phase to anhydrous sodium sulfate, Na_2SO_4, which dissolves with evolution of heat. This change is evidenced by increased solubility of the hydrated salt up to 32.4°, but above this temperature the solubility decreases.

These temperature effects are what would be predicted from Eq 4. When the heat of solution is negative, signifying that energy is released during dissolution, the relation between $\log K_S$ and $1/T$ is typified in Fig 16-3 (Curve *A*), where as $1/T$ increases, $\log K_S$ increases. It can be seen that with increasing temperature (T itself actually increases proceeding left in Fig 16-3, *A*) there is a decrease in solubility. On the other hand, when the heat of solution is positive—that is, when heat is absorbed in the solution process—the relation between $\log K_S$ and $1/T$ is typified in Fig 16-3, *B*. Here, as temperature increases ($1/T$ decreases), the solubility increases.

Effect of Salts—The solubility of a nonelectrolyte, in water, either is decreased or increased generally by the addition of an electrolyte; it is only rarely that the solubility is not altered. When the solubility of the nonelectrolyte is decreased, the effect is referred to as *salting-out*; if it is increased, it is described as *salting-in*. Inorganic electrolytes commonly decrease solubility, though there are some exceptions to the generalization.

Salting-out occurs because the ions of the added electrolyte interact with water molecules and, thus, in a sense reduce the amount of water available for solution of the nonelectrolyte. (Refer to the section on *Thermodynamics of the Solution Process*, page 215, for another view.) The greater the degree of hydration of the ions, the more the solubility of the nonelectrolyte is decreased. If, for example, one compares the effect of equivalent amounts of lithium chloride, sodium chloride, potassium chloride, rubidium chloride and cesium chloride (all of which belong to the family of alkali metals and are of the same valence type), it is observed that lithium chloride decreases the solubility of a nonelectrolyte to the greatest extent and that the salting-out effect decreases in the order given. This is also the order of the degree of hydration of the cations; lithium ion, being the smallest ion and, therefore, having the greatest density of positive charge per unit of surface area (see also Chapter 13 under *Electronegativity Values*), is the most extensively

hydrated of the cations while cesium ion is hydrated the least. Salting-out is encountered frequently in pharmaceutical operations.

Salting-in, commonly occurs when either the salts of various organic acids or organic-substituted ammonium salts are added to aqueous solutions of nonelectrolytes. In the first case the solubilizing effect is associated with the anion and in the second, with the cation. In both cases the solubility increases as the concentration of added salt is increased. The solubility increase may be relatively great, sometimes amounting to several times the solubility of the nonelectrolyte in water.

Solubility of Solutes Containing Two or More Species—In cases where the solute phase consists of two or more species (as in an ionizable inorganic salt), when the solute goes into solution, the solution phase often contains each of these species as discrete entities. For some such substance, *AB*, the following relationship for the solution process may be written.

$$AB_{(solid)} \rightleftharpoons A_{(solution)} + B_{(solution)}$$

Since there is an equilibrium between the solute and saturated solution phases, the Law of Mass Action defines an equilibrium constant, K_{eq}

$$K_{eq} = \frac{a_{A(solution)} \cdot a_{B(solution)}}{a_{AB(solid)}} \quad (6)$$

where $a_{A(solution)}$, $a_{B(solution)}$ and $a_{AB(solid)}$ are the activities of *A* and *B* in solution and of *AB* in the solid phase. Recall from the earlier discussion that the activity of a solid is defined as unity, and that in a very dilute solution (eg, for a slightly soluble salt), concentrations may be substituted for activities and Eq 6 then becomes

$$K_{eq} = C_A C_B$$

where C_A and C_B are the concentrations of *A* and *B* in solution. In this situation K_{eq} has a special name, the *solubility product*, K_{SP}. Thus

$$K_{SP} = C_A C_B \quad (7)$$

This equation will hold true theoretically only for slightly soluble salts.

As an example of this type of solution, consider the solubility of silver chloride

$$K_{SP} = [Ag^+][Cl^-]$$

where the brackets [ ] designate concentrations.

At 25° the solubility product has a value of 1.56×10^{-10}, the concentration of silver and chloride ions being expressed in moles/liter. The same numerical value applies also to solutions of silver chloride containing an excess of either silver or chloride ions. If the silver-ion concentration is increased by the addition of a soluble silver salt, the chloride-ion concentration must decrease until the product of the two concentrations again is equal numerically to the solubility product. In order to effect the decrease in chloride-ion concentration, silver chloride is precipitated and, hence, its solubility is decreased. In a similar manner an increase in chloride-ion concentration by the addition of a soluble chloride effects a decrease in the silver-ion concentration until the numerical value of the solubility product is attained. Again, this decrease in silver-ion concentration is brought about by the precipitation of silver chloride.

The solubility of silver chloride in a saturated aqueous solution of the salt may be calculated by assuming that the concentration of silver ion is the same as the concentration of chloride ion, both expressed in moles/liter, and that the concentration of dissolved silver chloride is numerically the same since each silver chloride molecule gives rise to one silver ion and one chloride ion. Since

$$[\text{dissolved AgCl}] = [Ag^+] = [Cl^-]$$

the solubility of AgCl is equal to $\sqrt{1.56 \times 10^{-10}}$, which is 1.25×10^{-5} mole/liter. Multiplying this by the molecular weight of silver chloride (143) we obtain a solubility of approximately 1.8 mg/liter.

For a salt of the type $PbCl_2$ the solubility product expression takes the form

$$[Pb^{2+}][Cl^-]^2 = K_{SP}$$

while for As_2S_3 it would be

$$[As^{3+}]^2[S^{2-}]^3 = K_{SP}$$

because from the Law of Mass Action

$$PbCl_{2(solid)} \rightleftharpoons Pb^{2+}{}_{(solution)} + 2Cl^-{}_{(solution)}$$

and

$$As_2S_{3(solid)} \rightleftharpoons 2As^{3+}{}_{(solution)} + 3S^{2-}{}_{(solution)}$$

For further details of methods of using solubility-product calculations, the reader is referred to books on qualitative or quantitative analysis or physical chemistry.

Recall that the solubility-product principle is valid for aqueous solutions of slightly soluble salts, provided the concentration of added salt is not too great. Where the concentrations are high, deviations from the theory occur and these have been explained by assuming that in such solutions the nature of the solvent has been changed. Frequently, deviations also may occur as the result of the formation of complexes between the two salts. An example of increased solubility, by virtue of complex-ion formation, is seen in the effect of solutions of soluble iodides on mercuric iodide. According to the solubility-product principle it might be expected that soluble iodides would decrease the solubility of mercuric iodide, but because of the formation of the more soluble complex salt K_2HgI_4 which dissociates as follows

$$K_2HgI_4 \rightleftharpoons 2K^+ + (HgI_4)^{2-}$$

the iodide ion no longer functions as a common ion.

Practical applications of the solubility-product principle are found in qualitative and quantitative analysis whenever an excess of a precipitant is added in order to diminish, by common-ion effect, the solubility of the precipitate.

It is possible to formulate some general rules regarding the effect of the addition of soluble salts to slightly soluble salts where the added salt does not have an ion common to the slightly soluble salt. If the ions of the added soluble salt are not highly hydrated (see *Effect of Salts on the Solubility of Nonelectrolytes*, page 209), the solubility product of the slightly soluble salt will increase because the ions of the added salt tend to decrease the interionic attraction between the ions of the slightly soluble salt. On the other hand, if the ions of the added soluble salt are hydrated, water molecules become less available and the interionic attraction between the ions of the slightly soluble salt increases with a resultant decrease in solubility product. Another way of considering this effect is discussed later (*Thermodynamics of the Solution Process*, page 215).

The effect of temperature is, in general, what would be expected; increasing the temperature of the solution results in an increase of the solubility product.

Solubility Following a Chemical Reaction—Thus far in this chapter the discussion has been concerned with solubility that comes about because of interplay of entirely physical forces. The dissolution of some substance resulted from overcoming the physical interactions between solute mole-

cules and solvent molecules by the energy produced when a solute molecule interacted physically with a solvent molecule. The solution process, however, can be facilitated also by a chemical reaction. Almost always the chemical enhancement of solubility in aqueous systems is due to the formation of a salt following an acid–base reaction.

An alkaloidal base, or any other nitrogenous base of relatively high molecular weight, generally is slightly soluble in water, but if the pH of the medium is reduced by addition of acid, the solubility of the base is increased, considerably so, as the pH continues to be reduced. The reason for this increase in solubility is that the base is converted to a salt, which is relatively soluble in water. Conversely, the solubility of a salt of an alkaloid or other nitrogenous base is reduced as pH is increased by addition of alkali.

The solubility of slightly soluble acid substances is, on the other hand, increased as the pH is increased by addition of alkali, the reason again being that a salt, relatively soluble in water, is formed. Examples of acid substances whose solubility is thus increased are aspirin, theophylline and the penicillins, cephalosporins and barbiturates. Conversely, the soluhility of salts of the same substances is decreased as the pH decreases.

Among some inorganic compounds a somewhat similar hehavior is observed. Tribasic calcium phosphate, $Ca_3(PO_4)_2$, for example, is almost insoluble in water, but if an acid is added its solubility increases rapidly with a decrease in pH. Tbis is because hydrogen ions have such a strong affinity for phosphate ions forming nonionized phosphoric acid, that the calcium phosphate is dissolved in order to release phosphate ions. Or, stated in another way, the solubilization is an example of a reaction in which a strong acid (the source of the hydrogen ions) displaces a weak acid.

In all of these examples soluhilization occurs as tbe result of an interaction of the solute with an acid or a base and that the species in solution is *not* the same as the undissolved solute. Compounds which do not react with either acids or bases are slightly, or not at all, influenced in their aqueous solubility by variations of pH. Such effects as may he observed are generally due to ionic *salt effects*.

It is possible to analyze quantitatively the solubility following an acid–base reaction by considering it as a two-step process. The first example is an organic acid, designated as *HA*, that is relatively insoluble in water. Its two-step dissolution can be represented as

$$HA_{(solid)} \rightleftarrows HA_{(solution)}$$

followed by

$$HA_{(solution)} \rightleftarrows H^+_{(solution)} + A^-_{(solution)}$$

The equilibrium constant for the first step is the solubility of *HA* ($K_S = [HA]_{solution}$), just as was developed earlier when no chemical reaction took place, and the equilibrium constant for the second step is the dissociation constant of the acid is

$$K_a = \frac{[H^+][A^-]}{[HA]}$$

Since the total amount of compound *in solution* is the sum of nonionized and ionized forms of the acid, the total solubility may be designated as $S_{t(HA)}$, or

$$S_{t(HA)} = [HA] + [A^-] = [HA] + K_a \frac{[HA]}{[H^+]} \qquad (8)$$

Since $K_S = [HA]$, Eq 8 becomes

$$S_{t(HA)} = K_S\left(1 + \frac{K_a}{[H^+]}\right) \qquad (9)$$

Equation 9 is very useful since it equates the total solubility of an acid drug with the hydrogen-ion concentration of the solvent. If the water solubility, K_S, and the dissociation constant, K_a, are known, the total solubility of the acid can be calculated at various hydrogen-ion concentrations. Equation 9 demonstrates quantitatively how the total solubility of the acid increases as the hydrogen-ion concentration decreases (that is, as the pH increases).

It is possible to develop an equation similar to Eq 9 for the solubility of a basic drug, *B*, such as a relatively insoluble nitrogenous base (an alkaloid, for example), at various hydrogen-ion concentrations. The solubility of the base in water may be represented in two steps, as

$$B_{(solid)} \rightleftarrows B_{(solution)}$$

$$B_{(solution)} \rightleftarrows BH^+_{(solution)} + OH^-_{(solution)}$$

Again, if K_S is the solubility of the free hase in water and K_b is its dissociation constant

$$K_b = \frac{[BH^+][OH^-]}{[B]}$$

the total solubility of the hase in water $S_{t(B)}$ is given by

$$S_{t(B)} = [B] + [BH^+] = [B] + \frac{K_b[B]}{[OH^-]} =$$

$$K_S\left(1 + \frac{K_b}{[OH^-]}\right) \qquad (10)$$

It is convenient to rewrite Eq 10 in terms of hydrogen-ion concentration by making use of the dissociation constant for water

$$K_W = [H^+][OH^-] = 1 \times 10^{-14}$$

Equation 10 tben becomes

$$S_{t(B)} = K_S\left(1 + \frac{K_b}{K_W/[H^+]}\right) = K_S\left(1 + \frac{K_b[H^+]}{K_W}\right) \qquad (11)$$

Equation 11 quantitatively shows how the total solubility of the base increases as the hydrogen-ion concentration of the solvent increases. If K_S and K_b are known, it is possihle to calculate the total solubility of a basic drug at various hydrogen-ion concentrations using this equation.

Equations 9 and 11 have assumed that the salt formed following a chemical reaction is infinitely soluble. This, of course, is not an acceptable assumption as suggested and demonstrated by Kramer and Flynn.[2] Rather, for an acidic or basic drug there should be a pH at which *maximum solubility* occurs where this soluhility remains the sum of the solution concentrations of the free and salt forms of the drug at that pH. Using a basic drug, B, as the example, this would mean that a solution of B, at pH values greater than the pH of maximum solubility, would be saturated with free-base form but not with the salt form and the use of Eq 11 would be valid for the prediction of solubility. On the other hand, at pH values less than the pH of maximum soluhility, the solution would be saturated with salt form and Eq 11 is no longer really valid. Since in this situation the total solubility of the base, $S_{t(B)}$, is

$$S_{t(B)} = [B] + [BH^+]_s$$

where the subscript, *s*, designates a solution saturated with salt, the correct equation to use at pH values less than the pH maximum would be

$$S_{t(B)} = [BH^+]_s\left(1 + \frac{[OH^-]}{K_b}\right) = [BH^+]_s\left(1 + \frac{K_W}{K_b[H^+]}\right) \qquad (12)$$

A relationship similar to Eq 12 likewise can be developed for an acidic drug at pHs greater than its pH of maximum solubility.

Effecting Solution of Solids in the Prescription Laboratory—The method usually employed by the pharmacist when soluble compounds are to be dissolved in water in compounding a prescription requires the use of the mortar and pestle. The ordinary practice is to crush the substance into fragments in the mortar with the pestle and pour the solvent on it, meanwhile stirring with the pestle until solution is effected. If definite quantities are used, and the whole of the solvent is required to dissolve the given weight of the salt, only a portion of the solvent should be added first and, when this is saturated, the solution is poured off and a fresh portion of solvent added. This operation is repeated until the solid is dissolved entirely and all the portions combined. Other methods of effecting solution are to shake the solid with the liquid in a bottle or flask or to apply heat to the substances in a suitable vessel. Substances vary greatly in the rate at which they dissolve; some are capable of producing a saturated solution quickly, others require several hours to attain saturation. All too often, in their haste to prepare a saturated solution, pharmacists fail to obtain the required degree of solution of solute.

With hygroscopic substances like pepsin, silver protein compounds and some others, the best method of effecting solution in water is to place the substance directly upon the surface of the water and then stir vigorously with a glass rod. If the ordinary procedure, such as using a mortar and pestle, is employed with these substances, gummy lumps are formed which are exceedingly difficult to dissolve.

The *solubility* of chemicals and the *miscibility* of liquids are important physical factors for the pharmacist to know, as they often have a bearing on intelligently and properly filling prescriptions. Mainly for the information of the pharmacist, the USP provides tabular data indicating the degree of solubility or miscibility of many official substances.

Determination of Solubility—For the pharmacist and pharmaceutical chemist the question of solubility is of paramount importance. Not only is it necessary to know solubilities when preparing and dispensing medicines, but such information is necessary to effect separation of substances in qualitative and quantitative analysis. Furthermore, the accurate determination of the solubility of a substance is one of the best methods for determining its purity.

The details of the determination of the solubility are affected markedly by the physical and chemical characteristics of the solute and solvent and also by the temperature at which the solubility is to be determined. Accordingly, it is not possible to describe a universally applicable method but, in general, the following rules must be observed in solubility determinations.

1. The purity of both the dissolved substance and the solvent is essential, since impurities in either affect the solubility.
2. A constancy of temperature must be maintained accurately during the course of the determination.
3. Complete saturation must be attained.
4. Accurate analysis of the saturated solution and correct expression of the results are imperative.

Consideration should be given also to the varying rates of solution of different compounds and to the marked effect of the degree of fineness of the particles on the time required for the saturation of the solution.

Many of the solubility data of USP have been determined with regard to the exacting requirements mentioned above.

Phase-Solubility Analysis—This procedure is one of the most useful and accurate methods for the determination of the purity of a substance. It involves the application of precise solubility methods to the principle that constancy of solubility, in the same manner as constancy of melting point, indicates that a material is pure or free from foreign admixture. It is important to recognize that the technique can be used to obtain the exact solubility of the pure substance without the necessity of the experimental material itself being pure.

The method is based on the thermodynamic principles of heterogeneous equilibria which are among the soundest of theoretical concepts of chemistry. Thus, it does not depend on any assumptions regarding kinetics or structure of matter, but is applicable to all species of molecules, and is sufficiently sensitive to distinguish between optical isomers. The requirements for an analysis are simple, since the equipment needed is basic to most laboratories and the quantities of substances required are small.

The standard solubility method consists of five steps:

1. Mixing, in separate systems, increasing amounts of a substance with measured amounts of a solvent.
2. Establishment of equilibrium for each system at identical constant temperature and pressure.
3. Separation of the solid phase from the solutions.
4. Determination of the concentration of the material dissolved in the various solutions.
5. Plotting the concentration of the dissolved material per unit of solvent (y-axis, or solution concentration) against the mass of material per unit of solvent (x-axis or system concentration).

The solubility method has been established on the sound theoretical principles of the Gibbs phase rule: $F = C - P + 2$, which relates C, the number of components, F, the degrees of freedom (pressure, temperature and concentration) and P, the number of phases for a heterogeneous equilibrium. Since solubility analyses are carried out at constant temperature and pressure, a pure solid in solution would show only one degree of freedom, because only one phase is present at concentrations below saturation. This is represented by section AB in Fig 16-4. For a pure solid in a saturated solution at equilibrium (Fig 16-4, BC), two phases are present, solid and solution; there is no variation in concentration and thus, at constant temperature and pressure, no degrees of freedom.

The curve ABC of Fig 16-4 represents the type of solubility diagram obtained for: (1) a pure material, (2) equal amounts of two or more materials having identical solubilities or (3) a mixture of two or more materials present in the

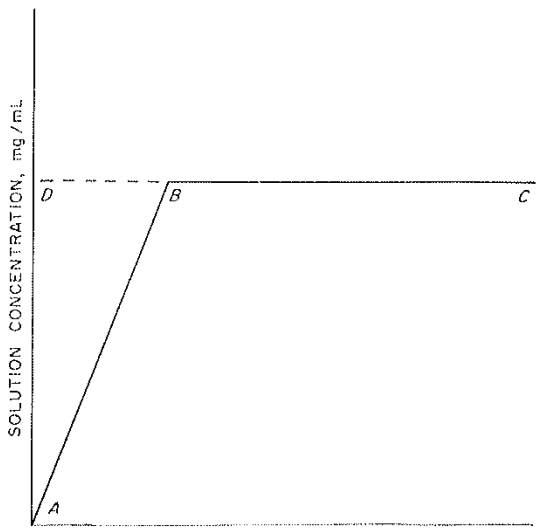


Fig 16-4. Phase-solubility diagram for a pure substance.

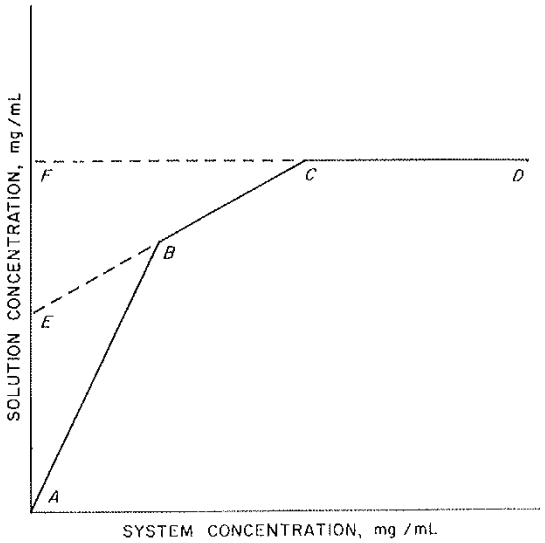


Fig 16-5. Type of solubility curve obtained when a substance contains one impurity.

unique ratio of their solubilities. These latter two cases are rare and often may be detected by a change in solvent system.

Line segment *BC* of Fig 16-4, since it has no slope, usually indicates purity. If, however, this section *does* exhibit a slope, its numerical value indicates the fraction of impurity present. Line segment *BC*, extrapolated to the *y* axis at *D*, is the actual solubility of the pure substance.

A representative type of solubility curve, which is obtained when a substance contains one impurity, is illustrated in Fig 16-5. Here, at *B*, the solution becomes saturated with one component. From *B* to *C* there are two phases present: a solution saturated with Component I (usually the major component) containing also some Component II (usually the minor component), and a solid phase of Component I. The one degree of freedom revealed by the slope of the line segment *BC* is the concentration of Component II, which is the impurity (usually the minor component). A mixture of *d* and *l* isomers could have such a curve, as would any simple mixtures in which the solubilities are independent of each other.

The section *CD* indicates that the solvent is saturated with both components of the two-component mixture. Here, three phases are present: a solution saturated with both components and the two solid phases. No variation of concentration is possible, hence, no degree of freedom is possible (indicated by the lack of slope of section *CD*). The distance *AE* on the ordinate represents the solubility of the major component, and the distance *EF*, the solubility of the minor component.

The fact that the equilibration process is time-consuming, requiring as long as 3 weeks in certain cases, is offset by the fact that all of the sample can be recovered after a determination. This adds to the general usefulness of the method, particularly in cases where the substance is expensive or difficult to obtain. A use for the method other than the determination of purity or of solubility is to obtain especially pure samples by recovering the solid residues at system concentration corresponding to points on section *BC* in Fig 16-5. Thus, the method is useful not only as a quantitative analytical tool, but also for purification.

Solutions of Liquids in Liquids

Binary Systems—Under this title the following types of liquid-pairs may be recognized.

1. Those which are soluble completely in each other in all proportions. Examples: alcohol and water; glycerin and water; alcohol and glycerin.
2. Those which are soluble in each other in definite proportions. Examples: phenol and water; ether and water; nicotine and water.
3. Those which are imperceptibly soluble in each other in any proportion. Examples: castor oil and water; liquid petrolatum and water.

The mutual solubility of liquid pairs of Type 2 has been studied extensively and found to show interesting regularities. If a series of tubes containing varying, but known, percentages of phenol and water are heated (or cooled, if necessary) just to the point of formation of a homogeneous solution, and the temperatures at such points noted, there will be obtained, upon plotting the results, a curve similar to that in Fig 16-6.[3] On this graph the area inside the curve represents the region where mixtures of phenol and water will separate into two layers, while in the region outside of the curve homogeneous solutions will be obtained. The maximum temperature on this curve is called the *critical solution temperature*, that is, the temperature above which a homogeneous solution occurs regardless of the composition of the mixture. For phenol and water the critical solution temperature occurs at a composition of 34.5% phenol in water.

Temperature versus composition curves, as depicted in Fig 16-6, provide much useful information in the preparation of homogeneous mixtures of substances showing mutual-solubility behavior. At room temperature (here assumed to be 25°), by drawing a line parallel to the abscissa at 25°, we find that we actually can prepare two sets of homogeneous solutions, one containing from 0 to about 7.5% phenol and the other containing phenol from 72 to about 95% (its limit of solubility). At compositions between 7.5 and 72% phenol at 25° two liquid layers or phases will separate. In sample tubes containing a concentration of phenol in this two-layer region at 25° one layer always will be phenol-rich and always contain 72% phenol while the other layer will be water-rich and always contain 7.5% phenol. These values are obtained by interpolation of the two points of intersection of the line drawn at 25° with the experimental curve. As it may be deduced, at other temperatures, the composition of the two layers in the two-layer region is determined by the points of intersection of the curve with a line (called

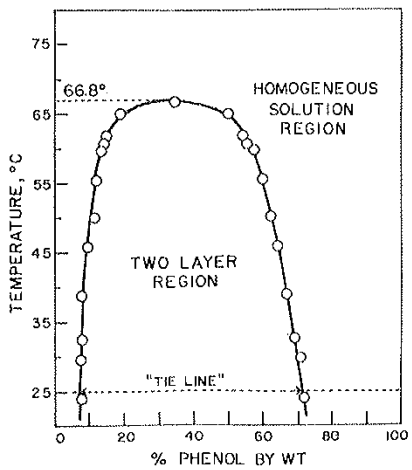


Fig 16-6. Phenol-water solubility.[3]

the "tie line") drawn parallel to the abscissa at that temperature. The relative amounts of the two layers or phases, phenol-rich and water-rich in this example, will depend on the concentration of phenol added. As expected, the proportion of phenol-rich layer relative to the water-rich layer increases as the concentration of phenol added increases. For example, at 20% phenol in water at 25° there would be more of the water-rich layer than of the phenol-rich layer, while at 50% phenol in water there would be more of the phenol-rich layer. The relative proportion of each layer may be calculated from such tie lines at any temperature and composition as well as the amount of phenol present in each of the two phases. To determine how these calculations are made and for further discussion of this topic the student should consult Ref 1, page 79.

A simple and practical advantage in the use of phase diagrams is pointed out in Ref 1. Based on diagrams such as Fig 16-6, they point out that the most concentrated stock solution of phenol that perhaps should be used by pharmacists is one containing 76% *w/w* phenol in water (equivalent to 80% *w/v*). At room temperature this mixture is a homogeneous solution and will remain homogeneous to around 3.5°, at which temperature freezing occurs. It should be noted that Liquefied Phenol USP contains 90% *w/w* phenol and freezes at 17°C. This means that if the storage area in the pharmacy falls to about 63°F, the preparation will freeze, resulting in a stock solution no longer convenient to use.

In the case of phenol and water the mutual solubility increases with increase in temperature and the critical solution temperature occurs at a relatively high point. In a certain number of cases, however, the mutual solubility increases with decrease in temperature and the critical solution temperature occurs at a relatively low value. Most of the substances which show lower critical solution temperatures are amines as, for example, triethylamine with water.

In addition to pairs of liquids which show *either* upper or lower critical solution temperatures, there are other pairs which show *both* upper and lower critical solution temperatures and the mutual solubility curve is of the closed type. An example of this type of liquid pair is found in the case of nicotine and water (Fig 16-7). Mixtures of nicotine and water represented by points within the curve will separate into two layers, but mixtures represented by points outside of the curve are perfectly miscible with each other.

In a discussion of solutions of liquids in liquids it is evident that the distinction between the terms solute and solvent loses its significance. For example, in a solution of water and glycerin, which shall be considered to be the solute and which the solvent? Again, when two liquids are soluble only partially in each other the distinction between solute and solvent might be reversed easily. In such cases the term solvent usually is given to the constituent present in larger quantity.

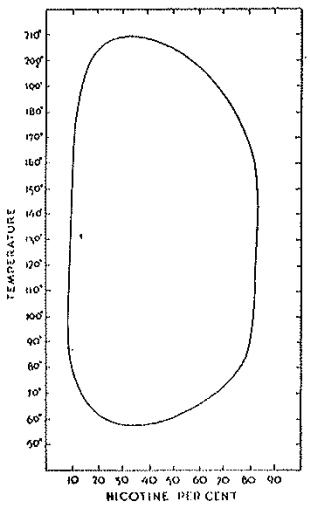


Fig 16-7. Nicotine-water solubility.

Ternary Systems—The addition of a third liquid to a binary liquid system to produce a ternary or three-component system can result in several possible combinations.

If the third liquid is soluble in only one of the two original liquids or if its solubility in the two original liquids is markedly different, the mutual solubility of the original pair will be decreased. An upper critical solution temperature will be elevated and a lower critical solution temperature lowered. On the other hand, the addition of a liquid having roughly the same solubility in both components of the original pair will result in an increase in their mutual solubility. An upper critical solution temperature then will be lowered and a lower critical solution temperature elevated.

An equilateral-triangle graph may be used to represent the situation in which a third liquid is added to a partially miscible liquid pair, the third liquid being miscible with each member of the original pair. In this type of graph, each side of the triangle represents 0% of one of the components and the apex opposite that side represents 100% of that component. The reader is referred to textbooks on experimental physical chemistry for details of the construction and use of graphs of this type.

Two other possibilities exist in ternary liquid systems: that in which two components are completely miscible and the third is partially miscible with each, and that in which all combinations of two of the three components are only partially miscible.

Solutions of Gases in Liquids

Nearly all gases are more or less soluble in liquids. One has but to recall the solubility of carbon dioxide, hydrogen sulfide or air in water as common examples.

The amount of gas dissolved in a liquid in general follows *Henry's law*, which states that *the weight of gas dissolved by a given amount of a liquid at a given temperature is proportional to its pressure*. Thus, if the pressure is doubled, twice as much gas will dissolve as at the initial pressure. The extent to which a gas is dissolved in a liquid, at a given temperature, may be expressed in terms of the *solubility coefficient*, which is the volume of gas measured under the conditions of the experiment, that is, absorbed by one volume of the liquid. The degree of solubility also is expressed sometimes in terms of the *absorption coefficient*, which is the volume of gas, reduced to standard conditions, dissolved by one volume of liquid under a pressure of one atmosphere.

Although Henry's law expresses fairly accurately the solubility of slightly soluble gases, it deviates considerably in the case of very soluble gases such as hydrogen chloride and ammonia. Such deviations most frequently are due to chemical interaction of solute and solvent.

The solubility of gases in liquids *decreases* with a *rise in temperature* and, in general, also when salts are added to the solvent, the latter effect being referred to as the *salting-out* of the gas.

Solutions of gases potentially are dangerous when exposed to warm temperatures because of the liberation and expansion of the dissolved gas which may cause the container to burst. Bottles containing such solutions (eg, strong ammonia solution) should be cooled before opening, if practical, and the stopper should be covered with a cloth before attempting its removal.

Thermodynamics of the Solution Process

In the discussions under this heading the solute is assumed to be in the liquid state, hence, the heat of solution ($\Delta H'$) is a term different from that in Eq 3 (ΔH). The heat of solution for a solid solute going into solution as defined in Eq 3 is the net heat effect for the overall dissolution

$$A_{(\text{solid})} \rightleftarrows A_{(\text{liquid})} \rightleftarrows A_{(\text{solution})}$$

Considering only the process

$$A_{(\text{liquid})} \rightleftarrows A_{(\text{solution})}$$

and assuming that the solute is a liquid (or a supercooled liquid in the case of a solid) at a temperature close to room temperature, where the energy needed for melting (heat of fusion) is not being considered.

For a physical or chemical reaction to occur spontaneously at a constant temperature and pressure, the net free-energy change, ΔG, for the reaction should be negative (see *Thermodynamics*, Chapter 15). Furthermore, it is known that the free-energy change depends on heat-related enthalpy ($\Delta H'$) and order-related entropy (ΔS) factors as seen in

$$\Delta G = \Delta H' - T\Delta S \tag{13}$$

where T is the temperature. Recall, also, that the relation between free energy and the equilibrium constant, K, for a reaction is given by

$$\Delta G = -RT \ln K \tag{14}$$

Equations 13 and 14 certainly apply to the solution of a drug. Since the solubility is, in reality, an equilibrium constant, Eq 14 indicates that the greater the negative value of ΔG, the greater the solubility.

The interplay of these two factors, $\Delta H'$ and ΔS in Eq 13, determines the free-energy change and, hence, whether or not dissolution of a drug will occur spontaneously. Thus, if in the solution process $\Delta H'$ is negative and ΔS positive, dissolution is favored since ΔG will be negative.

As the heat of solution is quite significant in the dissolution process one must look at its origin. (For an excellent and more complete discussion of the interactions and driving forces underlying the dissolution process, see Higuchi.[4]) The mechanism of solubility involves severing of the bonds that hold together the ions or molecules of a solute, the separation of molecules of solvent to create a space in the solvent into which the solute can be fitted and the ultimate response of solute and solvent to whatever forces of interaction may exist between them. In order to sever the bonds between molecules or ions of solute in the liquid state, energy must be supplied, as is the case also when molecules of solvent are to be separated. If heat is the source of energy it is apparent that both processes require the absorption of heat.

Solute–solvent interaction, on the other hand, generally is accompanied by the evolution of heat since the process occurs spontaneously. In effecting solution there is, accordingly, a heat-absorbing effect and a heat-releasing effect to be considered beyond those required to melt a solid. If there is no, or very little, interaction between solute and solvent, the only effect will be that of absorption of heat to produce the necessary separations of solute and solvent molecules or ions. If there is a significant interaction between solute and solvent, the amount of heat in excess of that required to overcome the solute–solute and the solvent–solvent forces is liberated. If the opposing heat effects are equal, there will be no change of temperature.

When $\Delta H'$ is zero, and there is no volume change, an *ideal solution* is said to exist since the solute–solute, solvent–solvent and solute–solvent interactions are the same. For such an ideal solution, the solubility of a solid can be predicted from its heat of fusion (the energy needed to melt the solid) at temperatures below its melting point. The student is referred to Ref 1, page 281 to see how this calculation is made.

When the heat of solution has a positive (energy absorbed) or negative (energy liberated) value, the solution is said to be a *nonideal solution*. A negative heat of solution favors solubility while a positive heat works against dissolution.

The magnitude of the various attractive forces involved between solute, solvent and solute–solvent molecules may vary greatly and thus could lead to varying degrees of positive or negative enthalpy changes in the solution process. The reason for this is that the molecular structure of the various solutes and solvents determining the interactions can themselves vary greatly. (For a discussion of these effects, see Ref 1, page 41.)

The solute–solute interaction that must be overcome can vary from the strong ion–ion interaction (as in a salt), to the weaker dipole–dipole interaction (as in nearly all organic medicinals that are not salts), to the weakest induced dipole-induced dipole interaction (as with naphthalene).

The attractive forces in the solvent that must be overcome are, most frequently, the dipole–dipole interaction (as found in water or acetone) and the induced dipole–induced dipole interaction (as in liquid petrolatum).

The energy-releasing solute–solvent interactions that must be taken into account may be one of four types. In decreasing energy of interaction these are ion–dipole interactions (eg, a sodium ion interacting with water), dipole–dipole interactions (eg, an organic acid dissolved in water), dipole–induced dipole interaction, to be discussed later (eg, an organic acid dissolved in carbon tetrachloride) and induced dipole–induced dipole interactions (eg, naphthalene dissolved in benzene).

Since the energy-releasing solute–solvent interaction should approximate the energy needed to overcome the solute-solute and solvent–solvent interactions, it should be apparent why it is not possible to dissolve a salt like sodium chloride in benzene. The interaction between the ions and benzene does not supply enough energy to overcome the interaction between the ions in the solute and therefore gives rise to a positive heat of solution. On the other hand, the interaction of sodium and chloride ions with water molecules does provide an amount of energy approximating the energy needed to separate the ions in the solute and the molecules in the solvent.

Consideration must next be given to entropy effects in dissolution processes. Entropy is an indicator of the disorder or randomness of a system. The more positive the entropy change (ΔS) is, the greater the degree of randomness or disorder of the reaction system and the more favorably disposed is the reaction. Unlike $\Delta H'$, the entropy change (an entropy of mixing) in an ideal solution, is not zero, but has some positive value since there is an increase in the disorderliness or entropy of the system upon dissolution. Thus, in an ideal solution with $\Delta H'$ zero and ΔS positive, ΔG would have a negative value and the process, therefore, would be spontaneous.

In a nonideal solution, on the other hand, where $\Delta H'$ is not zero, ΔS can be equal to, greater than or less than the entropy of mixing found for the ideal solution. A nonideal solution with an entropy of mixing equal to that of the ideal solution is called a "regular solution." These solutions usually occur with nonpolar or weakly polar solutes and solvents. Such solutions are accompanied by a positive enthalpy change, implying that the solute–solvent molecular interaction is less than the solute–solute and solvent–solvent molecular interactions. Regular solutions are amenable to rigorous physical chemical analysis which will not be covered

in this chapter but which can be found in outline form in Ref 1, page 282.

The possibility exists in a nonideal solution that the entropy change is greater than for an ideal solution. Such a solution occurs when there is an association among solute or solvent molecules. In essence, then, the dissolution process occurs when one begins at a relatively ordered (low entropy) state and progresses to a disorderly (high entropy) state. The overall entropy change is positive, greater than that of the ideal case, and favorable to dissolution. As may be expected, the enthalpy change in such a solution is positive since association in a solute or solvent must he overcome. The facilitated solubility of citric acid (an unsymmetrical molecule), as compared to inositol (a symmetrical molecule), may be explained on the basis of such a favorable entropy change.[4]

The solubility of citric acid is greater than that of inositol, yet, on the basis of their heats of solution, inositol should be more soluble. One may regard this phenomenon in another way. The reason for the higher solubility of citric acid is that although there is no hindrance in the transfer of a citric acid molecule as it goes from the solute to the solution phase, when the structurally unsymmetrical citric acid attempts to return to the solute phase from solution, it must assume an orientation that will allow ready interaction with polar groups already oriented. If it does not have the required orientation, it will not return readily to the solute, but it will remain in solution, thus bringing about a solubility larger than expected on the basis of heat of solution.

On the other hand, the structurally symmetrical inositol, as it leaves the solution phase, can interact with the solute phase without requiring a definite orientation; all orientations are equivalent. Hence, inositol can enter the solute phase without hindrance and, therefore, no facilitation of its solubility is observed.

In general, unsymmetrical molecules tend to be more soluble than symmetrical molecules.

Another type of nonideal solution occurs where there is an entropy change less than that expected of an ideal solution. Such nonideal behavior can occur with polar solutes and solvents. In a nonideal solution of this type there is significant interaction between solute and solvent. As may be expected, the enthalpy change ($\Delta H'$) in such a solution is negative and favors dissolution, but this effect is tempered by the unfavorable entropy change occurring at the same time. The reason for the lower-than-ideal entropy change can be visualized where the equilibrium system is more orderly and has a lower entropy than that expected for an ideal solution. The overall entropy change of solution, thus, would be less and not favorable to dissolution. One may rationalize the lower-than-expected solubility of lithium fluoride on the basis of this phenomenon. Compared with other alkali halides, it has a solubility lower than would be expected based solely on enthalpy changes. Because of the small size of ions in this salt there may be considerable ordering of water molecules in the solution. This effect must, of course, lead to a lowered entropy and an unfavorable effect on solubility. The effect of soluble salts on the solubility of nonelectrolytes (page 209) or slightly soluble salts (page 210) may be considered a result of an unfavorable entropy effect.

Pharmaceutical Solvents

The discussion will focus now on solvents available to pharmacists and, in particular, on their interactions and properties of these solvents. It is most important that the pharmacist obtain an understanding of the possible differences in solubility of a given solute in various solvents since he most often is called on to select a solvent which will dissolve the solute. A knowledge of the properties of solvents will allow the intelligent selection of suitable solvents.

Molecular Interactions

The solvent–solvent interaction is, in pharmaceutical solvents, always made up of a dipole–dipole interaction (Keesom Force) and an induced dipole–induced dipole interaction (London Force). It is important to keep in mind that both forces are always present; the contribution that each of these forces makes toward the overall attractive force depends on the structure of the solvent molecule. Some solvents have interactions which predominantly involve the Keesom Force (eg, water), while others are predominantly composed of the London Force (eg, chloroform); usually, both forces will be found.

Dipole–Dipole Forces—The unequal sharing of the electron pair between two atoms due to a difference in their electronegativity brings about a separation of the positive and negative centers of electricity in the molecule, causing it to become polarized; that is, to assume a partial ionic character. The molecule then is said to be a *permanent dipole* and the substance described as being a *polar compound*.

The greater the difference in the electronegativities of the constituent atoms, the greater the inequality of sharing of the electron pair, the greater the distance between the positive and negative centers of electricity in the molecule and the more polar the resulting molecule. As the character of the bonds are intermediate between those existing in nonpolar compounds and those occurring in ionic salts, it is to be expected that the properties of polar compounds should be intermediate between those of the two other classes. Such, in fact, generally is the case.

Coordinate covalent compounds all are very strongly polar because both of the electrons constituting the bonding pair have been contributed by a donor atom which, in effect, loses an electron and becomes positively charged, while the acceptor atom may be considered to gain an electron and become negatively charged.

While, in general, the electronegativities of different kinds of atoms are different, and the expectation is, therefore, that all molecules containing two or more different atoms will be polar, many such molecules actually are nonpolar. Thus, while the electronegativity of chlorine is appreciably different from that of carbon, the molecule of carbon tetrachloride, CCl_4, is nonpolar because the symmetrical arrangement of chlorine atoms about the carbon atom is such as to cancel the effects of the difference in the electronegativity of the constituent atoms. The same is true in the case of methane, CH_4, and for hydrocarbons generally. But the molecules CH_3Cl, CH_2Cl_2 and $CHCl_3$ definitely are polar because of the unsymmetrical distribution of the forces within the molecule.

A knowledge of the degree of polarity of various molecules usually is available in the measurement of the *dipole moment*, μ, of the molecules. This quantity is defined as the product of one of the charges on the molecule and the distance between the two average centers of positive and negative electricity. Measurements of the dipole moment of a substance are made, when possible, on the vapor of the substance but, when not possible, a dilute solution of the substance in a nonpolar solvent is employed. Table II lists the values of the dipole moment for a number of substances.

As stated previously, the molecules of nonpolar substances are characterized by weak attractions for one another, while molecules of polar substances exhibit a relatively strong attraction, which is all the more powerful the greater the dipole moment. The reason for this is readily apparent; the dipoles tend to align themselves so that the opposite charges of two different molecules are adjacent. They affect

Table II—Dipole Moments

Substance	Electrostatic Units ($\mu \times 10^{18}$)
Water	1.85
Acetone	2.8
Methyl alcohol	1.68
Ethyl alcohol	1.70
Phenol	1.70
Ethyl ether	1.14
Aniline	1.51
Nitrobenzene	4.19
o-Dinitrobenzene	6.0
m-Dinitrobenzene	3.8
p-Dinitrobenzene	0.3
Benzene	0
Methane	0
Chloromethane	1.86
Dichloromethane	1.58
Chloroform	1.05
Carbon tetrachloride	0
Carbon monoxide	0.11
Carbon dioxide	0
Oxygen	0
Hydrogen	0
Hydrogen chloride	1.03
Hydrogen bromide	0.78
Hydrogen iodide	0.38
Hydrogen sulfide	0.95
Hydrogen cyanide	2.93
Ammonia	1.49

each other in somewhat the same manner as do two bar magnets, the opposite poles of which are adjacent. While thermal agitation tends to break up the alignment or association of the dipoles, there is, nevertheless, a resultant significant intermolecular force present.

Induced Dipole–Induced Dipole Forces—It is of interest to inquire at this point what force does exist between the molecules of compounds which are nonpolar, eg, those which have zero dipole moment. If some attractive force did not exist, the molecules could not be expected to cling together, as in the solid and liquid states. Although the attraction is relatively slight, there is a force that arises from momentary polarization of the molecules because of electronic oscillations which are taking place continuously within the molecules. The *temporary dipoles* thus produced induce opposite polarizations in adjacent molecules and the net effect is that there is a small but definite attractive force between the molecules to keep them together in the liquid and solid states. This attraction resulting from mutual polarization commonly is referred to as the London Force and as an induced dipole–induced dipole force.

The Hydrogen Bond—The attraction between oppositely charged ends of two dipoles is accentuated when the positive end of one dipole contains a hydrogen atom and the negative end of the other dipole contains an atom of fluorine, oxygen or nitrogen. In such instances the nucleus of the hydrogen atom—which is a proton—appears to be able to bind together the negative end of the molecule, of which it is a part, with the negative end of the adjacent molecule. This may be represented by Fig 16-8.

Since the proton is the smallest positively charged atomic particle, it can draw together two negatively charged atoms

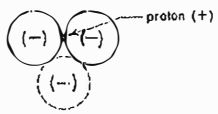
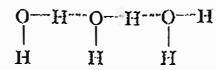


Fig 16-8. Hydrogen bonding.

or ions more closely than can any other—and necessarily larger—positively charged particle. Not more than two negative atoms are capable of being attracted at any given instant, as is evident from Fig 16-8, where a third negative atom is shown to be restricted physically from direct contact with the proton. Water is an excellent example of a substance, the molecules of which are associated through the formation of such a bond—called the *hydrogen bond*. An illustration of such bonding in the case of water may be represented as

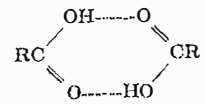


Each dotted line represents the bond or "bridge" established by the hydrogen atom of one water molecule with the oxygen atom of another. It is to be noted that the water molecule is pictured as an angular, rather than as a linear, molecule (H—O—H). This is in accord with the bond angles imposed by the directional character of the bonding orbitals making up the molecule (see Chapter 22). By virtue of its kernel containing six unneutralized protons, not only the valence electrons of the oxygen atom, but also those of the hydrogen atoms are attracted so strongly to the oxygen atom as to make the latter charged negatively, while the rest of the molecule is charged positively.

The hydrogen bond is not a strong bond, but it plays an important role in determining the properties of substances in which it occurs. For example, it primarily is responsible for the unusual properties of water. If the substance H_2O followed the course of the related substances H_2Te, H_2Se and H_2S, in so far as the physical properties of these latter substances are concerned, the freezing point of water would be about −100° and its boiling point about −80°. The unexpectedly high values actually observed are attributed to hydrogen bonding between molecules of water. To break such bonds, as for example in vaporizing water in the form of single H_2O molecules during the process of boiling, more energy is required than would be necessary if the water molecules were not linked by hydrogen bonds.

The molecules of at least the low-molecular-weight alcohols similarly are joined by hydrogen bonds to form a lattice-like structure.

Another example of the manner in which the hydrogen bond functions is seen in the case of carboxylic acids. Such acids usually exist in dimeric form, the two molecules being joined by hydrogen bonding, which may be depicted as



This tendency is so pronounced in the case of acetic acid that even in the vapor state the substance exists in dimeric aggregation.

Classification

On the basis of the forces of interaction occurring in solvents one may broadly classify solvents as one of three types:

1. *Polar solvents*—those made up of strong dipolar molecules having hydrogen bonding (water or hydrogen peroxide).
2. *Semipolar solvents*—those also made up of strong dipolar molecules but which do not form hydrogen bonds (acetone or pentyl alcohol).
3. *Nonpolar solvents*—those made up of molecules having a small or no dipolar character (benzene, vegetable oil or mineral oil).

Naturally, there are many solvents that may fit into more than one of these broad classes; for example, chloroform is a weak dipolar compound but generally is considered nonpo-

lar in character, and glycerin could be considered a polar or semipolar solvent even though it is capable of forming hydrogen bonds.

Types

Water—Water is a unique solvent. Besides being a highly associated liquid, giving rise to its high boiling point, it has another very important property, a high dielectric constant. The dielectric constant (ϵ) indicates the effect that a substance has, when it acts as a medium, on the ease with which two oppositely charged ions may be separated. The higher the dielectric constant of a medium, the easier it is to separate two oppositely charged species in that medium. The dielectric constants of a number of liquids are given in Table III. The values listed are relative to a vacuum which, by definition, has a dielectric constant of unity. According to Coulomb's Law the force of attraction (F) between two oppositely charged ions is

$$F = \frac{Z_1 Z_2}{\epsilon r^2} \tag{15}$$

where Z_1 and Z_2 are the charges on the ions, r is the distance separating the oppositely charged ions and ϵ is the dielectric constant of the medium. Equation 15 indicates that the force of interaction between the oppositely charged ions is proportional inversely to the dielectric constant of the medium. Thus, the interactive force between a sodium and chloride ion in water at a distance r would be $^1/_{80}$ that of the same ions in a vacuum separated the same distance. Looking at this example in another way, Coulomb's Law suggests that it is much easier to keep sodium and chloride ions apart in water than in a vacuum. Consider another example: the relative ease with which the ions of sodium chloride may be kept apart the same distance in water, as compared to olive oil, would occur in the ratio of 80/3.1; that is, it is 80/3.1 times easier to keep these ions apart in water than it is in olive oil. The ease of solubilizing salts in solvents like water and glycerin can be explained on the basis of their high dielectric constant. In general, also, the more polar the solvent, the greater its dielectric constant.

There is a very close relationship between dielectric constant and the two types of interactions found in all solvents; that is, the dipole–dipole interaction (Keesom) and the induced dipole–induced dipole interaction (London). The dielectric constant is related to these two forces through a quantity called *total molar polarization*, P, which is a measure of the relative ease with which a charge separation may be made within a molecule. The total molar polarization is given by

Table III—Dielectric Constants (at 20°)

Hydrogen cyanide	116
Water	80
Glycerin	46
Ethylene glycol	41
Methyl alcohol	33
Ethyl alcohol	25
n-Propyl alcohol	22
Acetone	21
Aniline	7.0
Chloroform	5.0
Castor oil	4.6
Ethyl ether	4.3
Octyl alcohol	3.4
Olive oil	3.1
Benzene	2.2
Turpentine oil	2.2
Carbon tetrachloride	2.2
Octane	1.9

$$P = \frac{\epsilon - 1}{\epsilon + 2} \cdot \frac{M}{D} \tag{16}$$

where ϵ is the dielectric constant of the substance, M is the molecular weight and D is the density. (For further details, see Ref 1, page 114.) Total molar polarization is in turn composed of two terms

$$P = P_\alpha + P_\mu = \frac{4}{3}\pi N\alpha + \frac{4}{3}\pi N\left(\frac{\mu^2}{3kT}\right) \tag{17}$$

where $P_\alpha = 4/3\ \pi N\alpha$ is the contribution due to induced polarization (the London contribution), and where

$$P_\mu = \frac{4}{3}\pi N\left(\frac{\mu^2}{3kT}\right)$$

is the contribution due to the permanent dipole (the Keesom contribution), N is Avogadro's Number, α is a constant called the polarizability (related to the induced dipole), μ is the dipole moment, k is the Boltzmann constant (1.38×10^{-16} erg/mole/deg) and T is the absolute temperature. Grouping all constant terms, it is possible to rewrite Eq 17 as

$$P = A + B/T$$

and substituting Eq 16, yields

$$\frac{\epsilon - 1}{\epsilon + 2} \cdot \frac{M}{D} = A + \frac{B}{T} \tag{18}$$

The first term on the right-hand side is the contribution to the dielectric constant of the London dispersions; it is not temperature-dependent. The second term on the right-hand side is the contribution to the dielectric constant of the Keesom dispersions. This latter contribution is temperature-dependent because the contribution from the permanent dipole depends on the dipoles aligning themselves, which tendency is opposed by thermal agitation. Thus, it is apparent from Eq 18 (and from common sense) that as temperature increases, the dielectric constant of dipolar solvents will tend to decrease.

Equation 18 also indicates that solvents which have large dipole moments tend to have large dielectric constants because of the contribution of the P_μ term (Eq 17). Water, which has a very large dielectric constant, is estimated to have $^2/_3$ of its molecular interaction due to a dipole–dipole interaction and $^1/_3$ due to the induced dipole–induced dipole interaction. On the other hand, compounds such as benzene, with a dipole moment of zero, will have small dielectric constants since the contribution by the P_μ term will drop out of Eq 18.

There is an important concept that should be considered which has been introduced to pharmaceutical systems.[5] It must be recognized that pharmacists frequently are concerned with dissolving relatively nonpolar drugs in aqueous or mixed polar aqueous solvents. To understand what may be happening in such cases, factors concerned with the entropic effects arising from interactions originating with the nonpolar solutes must be considered. Previously it had been noted that the favorable entropic effect on dissolution was due to the disruption of associations occurring among solute or solvent molecules. Now, consider the effects on solubility due to solute interactions in the solution phase. Since the solutes under discussion are relatively nonpolar, the interactions are of the London Force type or a *hydrophobic association*. This hydropbobic association in aqueous solutions may cause significant structuring of water with a resultant ordered or low-entropy system, unfavorable to solution. As is known, the solution of an essentially nonpolar molecule in water is not a favorable process. It should be stressed that this is due to not only an unfavorable enthalpy change but also an unfavorable entropy change generated by

water-structuring. Such an unfavorable entropy change is quite significant in the solution process. As an example of this effect, the aqueous solubility of a series of alkyl *p*-aminobenzoates shows a ten million-fold decrease in solubility in going from the 1-carbon analog to the 12-carbon analog. These findings demonstrate clearly the considerable effect that hydrophobic associations can have.

Alcohols—*Ethanol*, as a solvent, is next in importance to water. An advantage is that growth of microorganisms does not occur in solutions containing alcohol in a reasonable concentration.

Resins, volatile oils, alkaloids, glycosides, etc are dissolved by alcohol, while many therapeutically inert principles, such as gums, albumin and starch, are insoluble, which makes it more useful as a "selective" solvent. Mixtures of water and alcohol, in proportions varying to suit specific cases, are used extensively. They are often referred to as *hydroalcoholic* solvents.

Glycerin is an excellent solvent, although its range is not as extensive as that of water or alcohol. In higher concentrations it has preservative action. It dissolves the fixed alkalies, a large number of salts, vegetable acids, pepsin, tannin, some active principles of plants, etc, but it also dissolves gums, soluble carbohydrates, starch, etc. It is also of special value as a simple solvent, as in phenol glycerite, or where the major portion of the glycerin simply is added as a preservative and stabilizer of solutions that have been prepared with other solvents (see *Glycerines*, Chapter 84).

Propylene glycol, which has been used widely as a substitute for glycerin, is miscible with water, acetone or chloroform in all proportions. It is soluble in ether and will dissolve many essential oils but is immiscible with fixed oils. It is claimed to be as effective as ethyl alcohol in its power of inhibiting mold growth and fermentation.

Isopropyl alcohol possesses solvent properties similar to those of ethyl alcohol and is used instead of the latter in a number of pharmaceutical manufacturing operations. It has the advantage in that the commonly available product contains not over 1% of water, while ethyl alcohol contains about 5% water, often a disadvantage. Isopropyl alcohol is employed in some liniment and lotion formulations. It cannot be taken internally.

General Properties—Low-molecular-weight and polyhydroxy alcohols form associated structures through hydrogen bonds just as in water. When the carbon-atom content of an alcohol rises above five, generally only monomers then are present in the pure solvent. Although alcohols have high dielectric constants, compared to other types of solvents, they are small compared to water. As has been discussed, the solubility of salts in a solvent should be paralleled by its dielectric constant. That is, as the dielectric constant of a series of solvents increases, the probability of dissolving a salt in the solvent increases. This behavior is observed for the alcohols. Table IV, taken from Higuchi,[4] shows how the solubility of salts follows the dielectric constant of the alcohols.

As mentioned earlier, absolute alcohol rarely is used pharmaceutically. However, hydroalcoholic mixtures such as elixirs and spirits frequently are encountered. A very useful generalization is that the dielectric properties of a mixed solvent, such as water and alcohol, can be approximated as the weighted average of the properties of the pure components. Thus, a mixture of 60% alcohol (by weight) in water should have a dielectric constant approximated by

$$\epsilon_{(\text{mixture})} = 0.6(\epsilon_{(\text{alcohol})}) + 0.4(\epsilon_{(\text{water})})$$

$$\epsilon_{(\text{mixture})} = 0.6(25) + 0.4(80) = 47$$

The dielectric constant of 60% alcohol in water is found experimentally to be 43, which is in close agreement with that just calculated. The dielectric constant of glycerin is 46, close to the 60% alcohol mixture. One would, therefore, expect a salt like sodium chloride to have about the same solubility in glycerin as in 60% alcohol. The solubility of sodium chloride in glycerin is 8.3 g/100 g of solvent and in 60% alcohol about 6.3 g/100 g of solvent. This agreement would be even closer if comparisons were made on a volume rather than weight basis. At least qualitatively it can be said that the solubility of a salt in a solvent or a mixed solvent very closely follows the dielectric constant of the medium or, conversely, that the polarity of mixed solvents is paralleled by their dielectric constant, based on salt solubility.

Table IV—Solubilities of Potassium Iodide and Sodium Chloride in Several Alcohols and Acetone[a]

Solvent	g KI/ 100 g Solvent	g NaCl/ 100 g Solvent
Water	148	35.9
Glycerin	...	8.3 (20°)
Propylene Glycol	50	7.1 (30°)
Methanol	17	1.4
Acetone	2.9	...
Ethanol	1.88	0.065
1-Propanol	0.44	0.0124
2-Propanol	0.18	0.003
1-Butanol	0.20	0.005
1-Pentanol	0.089	0.0018

[a] All measurements are at 25°C unless otherwise indicated.

Although the dielectric constant is useful in interpreting the effect of mixed solvents on salt solubility, it cannot be applied properly to the effect of mixed solvents on the solubility of nonelectrolytes. It was seen earlier that unfavorable entropic effects can occur upon dissolution of relatively nonpolar nonelectrolytes in water. Such an effect due to hydrophobic association considerably affects solubility. Yalkowsky[5] studied the ability of cosolvent systems to increase the solubility of nonelectrolytes in polar solvents where the cosolvent system essentially brings about a reduction in structuring of solvent. Thus, by increasing, in a positive sense, the entropy of solution by using cosolvents, it was possible to increase the solubility of the nonpolar molecule. Using as an example the solubility of alkyl *p*-aminobenzoates in propylene glycol-water systems, Yalkowsky[5] reported that it is possible to increase the solubility of the nonelectrolyte by several orders of magnitude hy increasing the fraction of propylene glycol in the aqueous system. Sometimes, it is found that, as a good first approximation, the logarithm of the soluhility is related linearly to the fraction of propylene glycol added by

$$\log S_f = \log S_{f=0} + \epsilon f$$

where S_f is the solubility in the mixed aqueous system containing the volume fraction f of nonaqueous cosolvent, $S_{f=0}$ is the solubility in water and ϵ is a constant (not dielectric constant) characteristic of the system under study. Specifically, when a 50% solution of propylene glycol in water is used, there is a 1000-fold increase in solubility of dodecyl *p*-aminohenzoate, in comparison to pure water.

In a series of studies, Martin *et al*[6] have made attempts to predict solubility in mixed solvent systems through an extension of the "regular solution" theory. The equations are logarithmic in nature and can reduce in form to the equations of Yalkowsky.[5]

Acetone and Related Semipolar Materials—Even though acetone has a very high dipole moment (2.8×10^{-18} esu), as a pure solvent it does not form associated structures.

Fig 16-9. The charge separation in acetone.

This is evidenced by its low boiling point (57°) in comparison with the boiling point of the lower-molecular-weight water (100°) and ethanol (79°). The reason why it does not associate is because the positive charge in its dipole does not reside in a hydrogen atom (Fig 16-9), precluding the possibility of its forming a hydrogen bond. However, if some substance which is capable of forming hydrogen bonds, such as water or alcohol, is added to acetone, a very strong interaction through hydrogen bonding will occur (see *Mechanism of Solvent Action* below). Some substances which are semipolar and similar to acetone are aldehydes, low-molecular-weight esters, other ketones and nitro-containing compounds.

Nonpolar Solvents—This class of solvents includes fixed oils such as vegetable oil, petroleum ether (ligroin), carbon tetrachloride, benzene and chloroform. On a relative basis there is a wide range of polarity among these solvents; for example, benzene has no dipole moment while that of chloroform is 1.05×10^{-18} esu. But even the polarity of these compounds normally classified as nonpolar is still in line with the dielectric constant of the solvent. The relation between these quantities is seen best through a quantity called *molar refraction*. The molar refraction (or *refractivity*), R, of a compound is given by

$$R = \frac{n^2 - 1}{n^2 + 2} \cdot \frac{M}{D} \tag{19}$$

where n is the refractive index of the liquid, M is its molecular weight and D is its density. The similarity between Eq 19 and Eq 16 is to be noted and, indeed, in refractive index measurements using very long wavelengths of light, $n^2 = \epsilon$. Thus, molar refraction under these conditions approximates total molar polarization. Since, in the more nonpolar solvents there is generally no dipole moment, μ, total molar polarization reflects polarization due only to the induced dipoles possible. Thus

$$P_\alpha = \frac{n^2 - 1}{n^2 + 2} \cdot \frac{M}{D} = \frac{\epsilon - 1}{\epsilon + 2} \cdot \frac{M}{D} = \frac{4}{3} \pi N \alpha \tag{20}$$

It is evident from this that the refractive index of a nonpolar compound reflects its relative polarity. For example, the more-polar benzene ($\epsilon = 2.2$) has a higher refractive index, 1.501, than the less-polar hexane ($\epsilon = 1.9$), whose refractive index is 1.375.

It should be emphasized again that when a solvent (such as chloroform) has highly electronegative halogen atoms attached to a carbon atom also containing at least one hydrogen atom, such a solvent will be capable of forming strong hydrogen bonds with solutes which are polar in character. Thus, through the formation of hydrogen bonds such solvents will dissolve polar solutes. For example, it is possible to dissolve alkaloids in chloroform.

Mechanism of Solvent Action

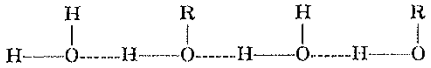
A solvent may function in one, or more, of several ways. When an ionic salt is dissolved, eg, by water, the process of solution involves separation of the cations and anions of the salt with attendant orientation of molecules of the solvent about the ions. Such orientation of solvent molecules about the ions of the solute—a process called *solvation* (*hydration*, if the solvent is water)—is possible only when the solvent is highly polar, whereby, the dipoles of the solvent are attracted to and held by the ions of the solute. The solvent also must possess the ability to keep the solvated, charged ions apart with minimal energy. The role of the dielectric constant in keeping this energy to a minimum has been discussed earlier.

A polar liquid such as water may exhibit solvent action also by virtue of its ability to break a covalent bond in the solute and bring about ionization of the latter. For example, hydrogen chloride dissolves in water and functions as an acid as a result of

$$HCl + H_2O \rightarrow H_3O^+ + Cl^-$$

The ions formed by this preliminary reaction of breaking the covalent bond subsequently are maintained in solution by the same mechanism as ionic salts.

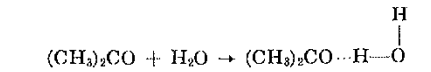
Still another mechanism by which a polar liquid may act as a solvent is that involved when the solvent and solute are capable of being coupled through hydrogen-bond formation. The solubility of the low-molecular-weight alcohols in water, for example, is attributed to the ability of the alcohol molecules to become part of a water-alcohol association complex.



As the molecular weight of the alcohol increases, it becomes progressively less polar and less able to compete with water molecules for a place in the lattice-like arrangement formed through hydrogen bonding; high-molecular-weight alcohols are, therefore, poorly soluble or insoluble in water. When the number of carbon atoms in a normal alcohol reaches five, its solubility in water is reduced materially.

When the number of hydroxyl groups in the alcohol is increased, its solubility in water generally is increased greatly; it is principally, if not entirely, for this reason that such high-molecular-weight compounds as sugars, gums, many glycosides and synthetic compounds, such as the polyethylene glycols, are very soluble in water.

The solubility of ethers, aldehydes, ketones, acids and anhydrides in water, and in other polar solvents, also is attributable largely to the formation of an association complex between solute and solvent by means of the hydrogen bond. The molecules of ethers, aldehydes and ketones, unlike those of alcohols, are not associated themselves, because of the absence of a hydrogen atom which is capable of forming the characteristic hydrogen bond. Notwithstanding, these substances are more or less polar because of the presence of a strongly electronegative oxygen atom, which is capable of association with water through hydrogen-bond formation. Acetone, for example, dissolves in water, in all likelihood, principally because of the following type of association:



The maximum number of carbon atoms which may be present per molecule possessing a hydrogen-bondable group, while still retaining water solubility, is approximately the same as for the alcohols.

Although nitrogen is less electronegative than oxygen and, thus, tends to form weaker hydrogen bonds, it is observed that amines are at least as soluble as alcohols containing an equivalent chain length. The reason for this is that alcohols form two hydrogen bonds with a net interaction of 12 kcal/mole. Primary amines can form three hydrogen bonds; two amine protons are shared with the oxygens of two water molecules, and the nitrogen accepts one water proton. The net interaction for the primary amine is between 12 and 13 kcal/mole and, hence, it shows an equal or greater solubility compared with corresponding alcohols.

The solvent action of nonpolar liquids involves a somewhat different mechanism. Because they are unable to form dipoles with which to overcome the attractions between ions of an ionic salt, or to break a covalent bond to produce an ionic compound or form association complexes with a solute, nonpolar liquids are incapable of dissolving polar compounds. They only can dissolve, in general, other nonpolar substances in which the bonds between molecules are weak. The forces involved usually are of the induced dipole–induced dipole type. Such is the case when one hydrocarbon is dissolved in another, or an oil or a fat is dissolved in petroleum ether. Sometimes it is observed that a polar substance, such as alcohol, will dissolve in a nonpolar liquid, such as benzene. This apparent exception to the preceding generalization may be explained by the assumption that the alcohol molecule induces a temporary dipole in the benzene molecule which forms an association complex with the solvent molecules. A binding force of this kind is referred to as a *permanent dipole–induced dipole force.*

Some Useful Generalizations—The preceding discussion indicates that enough is known about the mechanism of solubility to be able to formulate some generalizations concerning this important physical property of substances. Because of the greater importance of organic substances in the field of medicinal chemistry, certain of the more useful generalizations with respect to organic chemicals are presented here in summary form. It should be remembered, however, that the phenomenon of solubility usually involves several variables, and there may be exceptions to general rules.

One general maxim which holds true in most instances is: *the greater the structural similarity between solute and solvent, the greater the solubility.* As often stated to the student, "like dissolves like." Thus, phenol is almost insoluble in petroleum ether but is very soluble in glycerin.

Organic compounds containing polar groups capable of forming hydrogen bonds with water are soluble in water, providing that the molecular weight of the compound is not too great. It is demonstrated easily that the polar groups OH, CHO, COH, CHOH, CH_2OH, COOH, NO_2, CO, NH_2 and SO_3H tend to increase the solubility of an organic compound in water. On the other hand, nonpolar or very weak polar groups, such as the various hydrocarbon radicals, reduce solubility; the greater the number of carbon atoms in the radical, the greater the decrease in solubility. Introduction of halogen atoms into a molecule in general tends to decrease solubility because of an increased molecular weight without a proportionate increase in polarity.

Table V—Demonstration of Solubility Rules

Chemical Compound	Solubility[a]
Aniline, $C_6H_5NH_2$	28.6
Benzene, C_6H_6	1430
Benzoic acid, C_6H_5COOH	275
Benzyl alcohol, $C_6H_5CH_2OH$	25
1-Butanol, C_4H_9OH	12
t-Butyl alcohol, $(CH_3)_3COH$	Miscible
Carbon tetrachloride, CCl_4	2000
Chloroform, $CHCl_3$	200
Fumaric acid, *trans*-butenedioic acid	150
Hydroquinone, $C_6H_4(OH)_2$	14
Maleic acid, *cis*-butenedioic acid	5
Phenol, C_6H_5OH	15
Pyrocatechol, $C_6H_4(OH)_2$	2.3
Pyrogallol, $C_6H_3(OH)_3$	1.7
Resorcinol, $C_6H_4(OH)_2$	0.9

[a] The number of mL of water required to dissolve 1 g of solute.

The greater the number of polar groups contained per molecule, the greater the solubility of a compound, provided that the size of the rest of the molecule is not altered; thus, pyrogallol is much more soluble in water than phenol. The *relative positions* of the groups in the molecule also influence solubility; thus, in water, resorcinol (*m*-dihydroxybenzene) is more soluble than catechol (*o*-dihydroxybenzene), and the latter is more soluble than hydroquinone (*p*-dihydroxybenzene).

Polymers and compounds of high molecular weight generally are insoluble or only very slightly soluble.

High melting points frequently are indicative of low solubility for organic compounds. One reason for high melting points is the *association* of molecules and this cohesive force tends to prevent dispersion of the solute in the solvent.

The *cis* form of an isomer is more soluble than the *trans* form. See Table V.

Solvation, which is evidence of the existence of a strong attractive force between solute and solvent, enhances the solubility of the solute, provided there is not a marked ordering of the solvent molecules in the solution phase.

Acids, especially strong acids, usually produce water-soluble salts when reacted with nitrogen-containing organic bases.

Colligative Properties of Solutions

Up to this point concern has been with dissolving a solute in a solvent. Having brought about the dissolution, the solution, quite naturally, has a number of properties which are different from that of the pure solvent. Of very great importance are the *colligative properties* which a solution possesses.

The colligative properties of a solution are those that depend on the *number* of solute particles in solution, irrespective of whether these are molecules or ions, large or small. Ideally, the effect of a solute particle of one species is considered to be the same as that of an entirely different kind of particle, at least in dilute solution. Practically, there may be differences which may become substantial as the concentration of the solution is increased.

The colligative properties which will be considered are:

1. Osmotic pressure.
2. Vapor-pressure lowering.
3. Boiling-point elevation.
4. Freezing-point depression.

Of these four, all of which are related, osmotic pressure has the greatest direct importance in the pharmaceutical sciences. It is the property that largely determines the physiological acceptability of a variety of solutions used for therapeutic purposes.

Osmotic-Pressure Elevation

Diffusion in Liquids—Although the property of diffusion is rapid in gaseous systems, it is not limited to such systems. That molecules or ions in liquid systems possess this same freedom of movement may be demonstrated by placing carefully a layer of water on a concentrated aqueous solution of any salt. In time it will be observed that the boundary between solvent and solution widens gradually since salt moves into the water layer and water migrates from its layer into the salt solution below. Eventually, the composition of the new solution will become uniform throughout. This experiment indicates that *substances tend to move or diffuse from regions of higher concentration to regions of lower concentrations* so that differences in concentration eventually disappear.

Osmosis—In carrying out the experiment just described, it is impossible to distinguish between the diffusion of the solute and that of the solvent. However, by separating the solution and the solvent by means of a membrane that is permeable to the solvent, but not to the solute (such a membrane is referred to as a *semipermeable* membrane), it is possible to demonstrate visibly the diffusion of solvent into the concentrated solution, since volume changes will occur. In a similar manner, if two solutions of different concentra-

tion are separated by a membrane, the solvent will move from the solution of lower solute concentration to the solution of higher solute concentration. This diffusion of solvent through a membrane is called *osmosis*.

There is a difference between the activity or escaping tendency of the water molecules found in the solvent and salt solution separated by the semipermeable membrane. Since *activity*, which is related to water concentration, is higher on the pure solvent side, water moves from solvent to solution in order to equalize escaping-tendency differences. The difference in *escaping-tendency* gives rise to what is referred to as the *osmotic pressure* of the solution, which might be visualized as follows. A semipermeable membrane is placed over the end of a tube and a small amount of salt solution placed over the membrane in the tube. The tube then is immersed in a trough of pure water so that the upper level of the salt solution initially is at the same level as the water in the trough. With time, solvent molecules will move from solvent into the tube. The height of the solution will rise until the *hydrostatic pressure* exerted by the column of solution is equal to the *osmotic pressure*.

Osmotic Pressure of Nonelectrolytes—From quantitative studies with solutions of varying concentration of a solute that does not ionize, it has been demonstrated that *osmotic pressure is proportional to the concentration of the solute*; ie, twice the concentration of a given nonelectrolyte will produce twice the osmotic pressure in a given solvent. (This is not strictly true in solutions of fairly high solute concentration, but does hold quite well for dilute solutions.)

Furthermore, the osmotic pressures of solutions of different nonelectrolytes are proportional to the number of molecules in each solution. Stated in another manner, the osmotic pressures of two nonelectrolyte solutions of the same molal concentration are identical. Thus, a solution containing 34.2 g of sucrose (mol wt 342) in 1000 g of water has the same osmotic pressure as a solution containing 18.0 g of anhydrous dextrose (mol wt 180) in 1000 g of water. These solutions are said to be *isoosmotic* with each other because they have identical osmotic pressures.

A study of the results of osmotic-pressure measurements on different substances led the Dutch chemist Jacobus Henricus van't Hoff, in 1885, to suggest that the solute in a solution may be considered as being analogous to the molecules of a gas and the osmotic pressure as being produced by the bombardment of the semipermeable membrane by the molecules of solute. According to van't Hoff's theory the osmotic pressure of a solution is equal to the pressure which the dissolved substance would exert in the gaseous state if it occupied a volume equal to the volume of the solution. From this it follows that, just as in the case of a gas, there is a proportionality between pressure and concentration of dissolved substance. This proportionality is illustrated well by the values of the osmotic pressure of solutions of sucrose at 0° as determined by the Earl of Berkeley and EGJ Hartley and shown in Table VI.

In column PV of Table VI a quantitative confirmation, at least for fairly dilute solutions, of van't Hoff's oversimplified

Table VI—Osmotic Pressure of Sucrose Solutions

Conc. (g/L), C	Vol in L in Which 1 g Mole is Dissolved, V[a]	Pressure in Atmos P	P/C	PV
10.00	34.2	0.65	0.065	22.2
20.00	17.1	1.27	0.064	21.7
45.00	7.60	2.91	0.065	22.1
93.75	3.65	6.23	0.067	22.7

[a] These figures were obtained by calculating the volume of solution in which 342 (mol wt) g of sucrose would be dissolved.

though useful generalization is shown by the constancy of the values of the product PV. Recall that the product of the pressure and the volume of a gas, at constant temperature, is likewise constant (Boyle's law).

Van't Hoff also deduced that the osmotic pressure must be proportional to the absolute temperature, just as in Charles' law for gases, which deduction was confirmed by the experiments of several workers. From this it follows that the equation $PV = nRT$ is valid for dilute solutions of nonelectrolytes just as a similar equation is valid for gases. However, even as Boyle's law does not apply to gases under high pressures and at low temperatures, so van't Hoff's equation for osmotic pressure does not apply in concentrated solutions.

Osmotic Pressure of Electrolytes—In discussing the generalizations concerning the osmotic pressure of solutions of nonelectrolytes it was stated that the osmotic pressures of two solutions of the same molal concentration are identical. This generalization, however, cannot be made for solutions of electrolytes, ie, acids, alkalies and salts (see Chapter 17).

For example, sodium chloride is assumed to ionize as

$$NaCl \rightarrow Na^+ + Cl^-$$

It is evident that each molecule of sodium chloride that ionizes produces two ions and, if sodium chloride is completely ionized, there will be twice as many particles as would be the case if it were not ionized at all. Furthermore, if each ion has the same effect on osmotic pressure as a molecule, it might be expected that the osmotic pressure of the solution would be twice that of a solution containing the same molal concentration of nonionizing substance.

For solutions which yield more than two ions as, for example

$$K_2SO_4 \rightarrow 2K^+ + SO_4^{2-}$$

$$FeCl_3 \rightarrow Fe^{3+} + 3Cl^-$$

it is to be expected that the complete dissociation of the molecules would give rise to osmotic pressures that are three and four times, respectively, the pressure of solutions containing an equivalent quantity of a nonionized solute. Accordingly, the equation $PV = nRT$, which may be employed to calculate the osmotic pressure of a *dilute* solution of a nonelectrolyte, also may be applied to *dilute* solutions of electrolytes if it is changed to $PV = inRT$, where the value of i approaches the number of ions produced by the ionization of the strong electrolytes cited in the preceding examples. For weak electrolytes i represents the total number of particles, ions and molecules together, in the solution, divided by the number of molecules that would be present if the solute did not ionize. The experimental evidence indicates that in dilute solutions, at least, the osmotic pressures approach the predicted values. It should be emphasized, however, that in more concentrated solutions of electrolytes the deviations from this simple theory are considerable, due to interionic attraction, solvation and other factors.

Biological Aspects of Osmotic Pressure—Osmotic-pressure experiments were made as early as 1884 by the Dutch botanist Hugo de Vries in his study of plasmolysis, which term is applied to the contraction of the contents of plant cells placed in solutions of comparatively high osmotic pressure. The phenomenon is caused by the osmosis of water out of the cell through the practically semipermeable membrane surrounding the protoplasm. If suitable cells (eg, the epidermal cells of the leaf of *Tradescanta discolor*) are placed in a solution of higher osmotic pressure than that of the cell contents, water flows out of the cell, causing the contents to draw away from the cell wall. On the other hand, if the cells are placed in solutions of lower osmotic pressure, water enters the cell, producing an expansion

which is limited by the rigid cell wall. By immersing cells in a series of solutions of varying solute concentration, a solution may be found in which plasmolysis is barely detectable or absent. The osmotic pressure of such a solution is then the same, or very nearly the same, as that of the cell contents, and it is then said that the solution is *isotonic* with the cell contents. Solutions of greater concentration than this are said to be *hypertonic* and solutions of lower concentration, *hypotonic*.

Red blood cells, or erythrocytes, have been studied similarly by immersion into solutions of varying concentration of different solutes. When introduced into water or into sodium chloride solutions containing less than 0.90 g of solute per 100 mL, human erythrocytes swell, and often burst, because of the diffusion of water into the cell and the fact that the cell wall is not sufficiently strong to resist the pressure. This phenomenon is referred to as *hemolysis*. If the cells are placed in solutions containing more than 0.90 g of sodium chloride per 100 mL, they lose water and shrink. By immersing the cells in a solution containing exactly 0.90 g of sodium chloride in 100 mL, no change in the size of the cells is observed; since in this solution the cells maintain their "tone," the solution is said to be *isotonic* with human erythrocytes. For the reasons indicated it is desirable that solutions to be injected into the blood should be made isotonic with erythrocytes. The manner in which this may be done is described in Chapter 79.

Distinction between Isoosmotic and Isotonic—The terms isoosmotic and isotonic are not to be considered as equivalent, although a solution often may be described as being both isoosmotic and isotonic. If a plant or animal cell is in contact with a solution that has the same osmotic pressure as the cell contents, there will be no net gain or loss of water by either solution *provided* the cell membrane is impermeable to all the solutes present. Since the volume of the cell contents remains unchanged, the "tone", or normal state, of the cell is maintained, and the solution in contact with the cell may be described not only as being isoosmotic with the solution in the cell, but also as being isotonic with it. If, however, one or more of the solutes in contact with the membrane can pass through the latter, it is evident that the volume of the cell contents will change, thus altering the "tone" of the cell; in this case the two solutions may be isoosmotic, yet not be isotonic.

It is possible that some substances used in an injection dosage form can cause hemolysis of red blood cells, even when their concentrations are such as to produce solutions theoretically isoosmotic with the cells, because the solutes diffuse through the membrane of the cells. For example, a 1.8% solution of urea has the same osmotic pressure as a 0.9% solution of sodium chloride, but the former solution produces hemolysis of red blood cells; obviously the urea solution is not isotonic with the cells. To determine if a solution is isotonic with erythrocytes, it is necessary to determine the concentration of solute at which the cells retain their normal size and shape. A simple method for doing this was devised by Setnikar and Temelcou,[7] who determined the concentration of a solution at which red blood cells maintained a volume equal to that occupied in an isotonic solution of sodium chloride. The red cell volumes were determined by centrifuging suspensions of them in different solutions, using a hematocrit tube.

Vapor-Pressure Lowering

When a nonvolatile solute is dissolved in a liquid solvent the vapor pressure of the solvent is lowered. This easily can be described qualitatively by visualizing solvent molecules on the surface of the solvent, which normally could escape into the vapor, being replaced by solute molecules which have little if any vapor pressure of their own. For ideal solutions of nonelectrolytes the vapor pressure of the solution follows Raoult's law

$$P_A = X_A P_A^\circ \tag{21}$$

where P_A is the vapor pressure of the solution, P_A° is the vapor pressure of the pure solvent and X_A is the mole fraction of solvent. This relationship states that the vapor pressure of the solution is proportional to the number of molecules of solvent in the solution. Rearranging Eq 21 gives

$$\frac{P_A^\circ - P_A}{P_A^\circ} = (1 - X_A) = X_B \tag{22}$$

where X_B is the mole fraction of the solute. This equation states that the lowering of vapor pressure in the solution relative to the vapor pressure of the pure solvent—called simply the *relative vapor-pressure lowering*—is equal to the mole fraction of the solute. The *absolute* lowering of vapor pressure of the solution is defined by

$$P_A^\circ - P_A = X_B P_A^\circ \tag{23}$$

Example—Calculate the lowering of vapor pressure and the vapor pressure at 20°, of a solution containing 50 g of anhydrous dextrose (mol wt 180.16) in 1000 g of water (mol wt 18.02). The vapor pressure of water at 20°, in absence of air, is 17.535 mm.

First, calculate the lowering of vapor pressure, using Eq 23, in which X_B is the mole fraction of dextrose, defined by

$$X_B = \frac{n_B}{n_A + n_B}$$

where n_A is the number of moles of solvent and n_B is the number of moles of solute. Substituting numerical values

$$n_B = \frac{50}{180.2} = 0.278$$

$$n_B = \frac{1000}{18.02} = 55.5$$

$$X_B = \frac{0.278}{55.5 + 0.278} = 0.00498$$

the lowering of vapor pressure is

$$P_A^\circ - P_A = 0.00498 \times 17.535$$

$$= 0.0873 \text{ mm}$$

The vapor pressure of the solution is

$$P_A = 17.535 - 0.0873$$

$$= 17.448 \text{ mm}$$

Boiling-Point Elevation

In consequence of the fact that the vapor pressure of any solution of a nonvolatile solute is less than that of the solvent, the boiling point of the solution—the temperature at which the vapor pressure is equal to the applied pressure (commonly 760 mm)—must be higher than that of the solvent. This is clearly evident in Fig 16-10.

The relationship between the elevation of boiling point and the concentration of nonvolatile, nonelectrolyte solute may be derived from the Clausius–Clapeyron equation (see Chapter 15), which is

$$\frac{dP}{dT} = \frac{P \cdot \Delta H_{vap}}{RT^2} \tag{24}$$

Replacing the differential expression dP/dT by $\Delta P/\Delta T_b$, where ΔP is the lowering of vapor pressure and ΔT_b is the

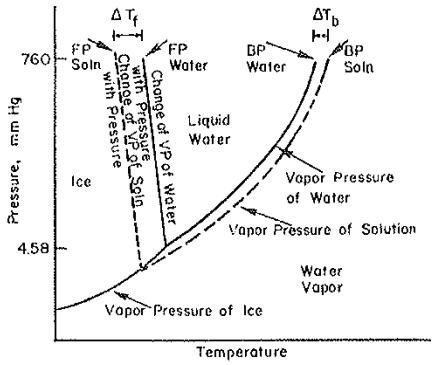


Fig 16-10. Vapor-pressure-temperature diagram for water and an aqueous solution, illustrating elevation of boiling point and lowering of freezing point of the latter.

elevation of boiling point, and introducing $P_A°$, the vapor pressure of the solvent at its boiling point T_0, results in

$$\frac{\Delta P}{\Delta T_b} = \frac{P_A° \cdot \Delta H_{vap}}{RT_0^2} \tag{25}$$

Since the lowering of vapor pressure in an ideal solution is

$$\Delta P = X_B P_A° \tag{26}$$

substitution of this equation into Eq 25, with rearrangement to provide a solution for ΔT_b, gives

$$\Delta T_b = \frac{RT_0^2}{\Delta H_{vap}} X_B \tag{27}$$

This equation may be used to calculate the elevation of the boiling point if the concentration of solute is expressed as the mole fraction. A more common expression, however, is in terms of the molality m (the number of gram-moles of solute per 1000 g of solvent), which relationship is derived as

$$X_B = \frac{n_B}{n_A + n_B} = \frac{m}{1000/M_A + m} \cong \frac{m}{1000/M_A} \tag{28}$$

In these equations M_A is the molecular weight of the solvent. When the solutions are dilute, so that m is small, it may be neglected in the denominator (*but not in the numerator!*) to give the approximate equivalent in Eq 28. Substituting this equivalent into Eq 27 gives

$$\Delta T_b = \frac{RT_0^2 M_A m}{1000 \Delta H_{vap}} \tag{29}$$

Grouping the constants into a single term results in

$$\Delta T_b = K_b m \tag{30}$$

where

$$K_b = \frac{RT_0^2 M_A}{1000 \Delta H_{vap}} \tag{31}$$

and is called the molal boiling-point elevation constant.

The value of this constant for water, which boils at 373.1° K, has a heat of vaporization of 539.7 cal/g and a molecular weight of 18.02, is

$$K_b = \frac{1.987 \times 373.1^2 \times 18.02}{1000 \times 18.02 \times 539.7} = 0.513° \tag{32}$$

Notwithstanding that K_b is called a molal boiling-point elevation constant, it should not be interpreted as the actual rise of boiling point for a 1-molal solution. Such solutions are generally too concentrated to exhibit the ideal behavior assumed in deriving the equation for calculating the theoretical value of the constant. In dilute solutions, however, the actual boiling-point elevation, *calculated to a 1-molal basis*, approaches the theoretical value, the closer the more dilute the solution.

The elevation of boiling point of a dilute solution of a nonelectrolyte solute may be used to calculate the mol wt of the latter. In a solution containing w_B g of solute of M_B in w_A g of solvent the molality m is

$$m = \frac{1000\, w_B}{w_A M_B} \tag{33}$$

substituting this into Eq 30 and rearranging gives

$$M_B = \frac{K_b 1000\, w_B}{w_A \Delta T_b} \tag{34}$$

Freezing-Point Depression

The freezing point of a solvent is defined as the temperature at which the solid and liquid forms of the solvent coexist in equilibrium at a fixed external pressure, commonly 1 atm (760 mm of mercury). At this temperature the solid and liquid forms of the solvent must have the same vapor pressure, for if this were not so, the form having the higher vapor pressure would change into that having the lower vapor pressure.

The freezing point of a solution is the temperature at which the solid form of the pure solvent coexists in equilibrium with the solution at a fixed external pressure, again commonly 1 atm. Since the vapor pressure of a solution is lower than that of its solvent, it is obvious that solid solvent and solution cannot coexist at the same temperature as solid solvent and liquid solvent; only at some lower temperature, where solid solvent and solution do have the same vapor pressure, is equilibrium established. A schematic pressure–temperature diagram for water and an aqueous solution, not drawn to scale and exaggerated for the purpose of more effective illustration, shows the equilibrium conditions involved in both freezing-point depression and boiling-point elevation (Fig 16-10).

The freezing-point lowering of a solution may be quantitatively predicted for ideal solutions, or dilute solutions which obey Raoult's law, by mathematical operations similar to (though somewhat more complex than) those used in deriving the boiling-point elevation constant. The equation for the freezing point lowering, ΔT_f, is

$$\Delta T_f = \frac{RT_0^2 M_A m}{1000 \Delta H_{fus}} = K_f m \tag{35}$$

where

$$K_f = \frac{RT_0^2 M_A}{1000 \Delta H_{fus}} \tag{36}$$

The value of K_f for water, which freezes at 273.1°K and has a heat of fusion of 79.7 cal/g, is

$$K_f = \frac{1.987 \times 273.1^2 \times 18.02}{1000 \times 18.02 \times 79.7} = 1.86° \tag{37}$$

The molal freezing-point depression constant is not intended to represent the freezing-point depression for a 1-molal solution, which is too concentrated for the premise of ideal behavior to be applicable. In dilute solutions the freezing-point depression, calculated to a 1-molal basis, approaches the theoretical value, the agreement between experiment and theory being the better the more dilute the solution.

The freezing point of a dilute solution of a nonelectrolyte solute may be used, as was the boiling point, to calculate the molecular weight of the solute. The applicable equation is

$$M_B = \frac{K_f 1000\, w_B}{w_A \Delta T_f} \tag{38}$$

The molecular weight of organic substances soluble in molten camphor may be determined by observing the freezing point of a mixture of the substance with camphor. This procedure, called the *Rast method*, uses camphor because it has a very large molal freezing-point-depression constant, about 40°. Since the "constant" may vary with different lots of camphor and with variations of technique, the method should be standardized using a solute of known molecular weight.

Freezing-point determinations of molecular weights have the advantage over boiling-point determinations of greater accuracy and precision by virtue of the larger magnitude of the freezing-point depression compared to boiling-point elevation. Thus, in the case of water the molal freezing-point depression is approximately 3.5 times greater than the molal boiling-point elevation.

Relationship between Osmotic Pressure and Vapor-Pressure Depression

The lowering of vapor pressure and the development of osmotic pressure in a solution are both manifestations of the basic condition that the free energy of solvent molecules in the pure solvent is greater than the free energy of solvent molecules in the solution. Consequently, solvent molecules will transfer spontaneously, if given an opportunity, from solvent to solution until equilibrium conditions are established. The transfer can take place either through a membrane permeable only to solvent molecules or, if such contact between solvent and solution is not available, by distillation of solvent from pure solvent to solution, if access through a vapor phase is provided.

If an experiment is performed with two sets of vessels containing solution and solvent, as illustrated in Fig 16-11, differing only in that the long tube of one set has a semipermeable membrane attached to its lower end, while in the other a hypothetical membrane separates the vapor phases, in time the same hydrostatic pressure should develop in both cases. For a definite volume, eg, a mole, of solvent transferred to the solution by distillation the change of free energy, ΔG, in the process is

$$\Delta G = RT \ln \frac{P_A}{P_A^\circ} \tag{39}$$

where P_A is the vapor pressure of the solution and P_A° is the vapor pressure of the solvent.

For the transfer of the same volume of solvent by osmosis the free-energy change is

$$\Delta G = -\bar{V}_A \pi \tag{40}$$

where $\bar{V}_A$ is the partial molal volume of solvent (the volume of 1 mole of solvent in the solution) and π is the osmotic pressure of the solution. Since the free-energy change is the same in both processes

$$-V_A \pi = RT \ln \frac{P_A}{P_A^\circ} \tag{41}$$

rearranging the equation yields

$$\pi = \frac{RT}{\bar{V}_A} \ln \frac{P_A^\circ}{P_A} \tag{42}$$

With this equation the osmotic pressure of a solution may be calculated if its vapor pressure and the partial molal volume of the solvent are known, not only when the solution is sufficiently dilute that Raoult's law is obeyed but also when the concentration is so high as to introduce substantial deviation from the law.

From Eq 42, which has some resemblance to van't Hoff's empirical equation $\pi V = nRT$ for dilute solutions, the latter equation may be derived as follows. If a solution is sufficiently dilute to correspond to Raoult's law, then

$$P_A = X_A P_A^\circ = (1 - X_B) P_A^\circ \tag{43}$$

and then Eq 42 may be written

$$\pi = -\frac{RT}{\bar{V}_A} \ln (1 - X_B) \tag{44}$$

When X_B is small (as in a dilute solution), the term $-\ln (1 - X_B)$ can be shown to be approximately equal to X_B, so that

$$\pi = \frac{RT}{\bar{V}_A} X_B \tag{45}$$

In dilute solutions the approximations $X_B = n_B/n_A$ (where n_B and n_A are the moles of solute and solvent, respectively) and $\bar{V}_A = V/n_A$ (where V is the volume of solution) may be introduced, yielding

$$\pi V = n_B RT \tag{46}$$

which is van't Hoff's equation.

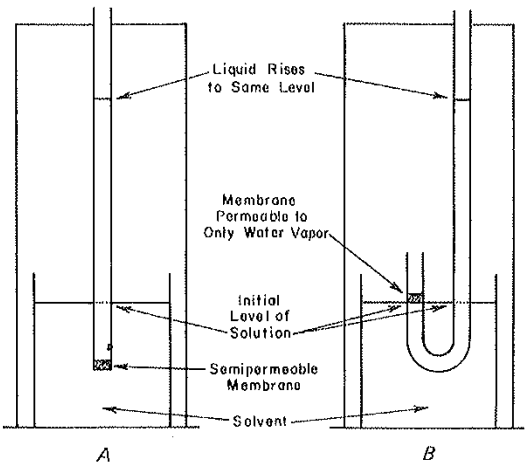


Fig 16-11. Transfer of solvent to equal volumes of solution. *A*: Osmotically, through a semipermeable membrane separating solution and solvent. *B*: By distillation, through a membrane separating solution and solvent vapor.

Ideal Behavior and Deviations

In setting out to derive mathematical expressions for colligative properties, such phrases as "for ideal solutions" or "for dilute solutions" were used to indicate the limitations of the expressions. Samuel Glasstone defines an ideal solution as "one which obeys Raoult's law over the whole range of concentration and at all temperatures" and gives as specific characteristics of such solutions their formation only from constituents which mix in the liquid state without heat change and without volume change. These characteristics reflect the fact that addition of a solute to a solvent produces no change in the forces between molecules of the solvent.

Thus, the molecules have the same *escaping-tendency* in the solution as in the pure solvent and the vapor pressure above the solution is proportional to the ratio of the number of solvent molecules in the surface of the solution to the number of the molecules in the surface of the solvent—which is the basis for Raoult's law.

Any change in intermolecular forces produced by mixing the components of a solution may result in deviation from ideality; such a deviation may be expected particularly in solutions containing both a polar and a nonpolar substance. Solutions of electrolytes, except at high dilution, are especially prone to depart from ideal behavior, even though allowance is made for the additional particles that result from ionization. When solute and solvent combine to form solvates, the escaping-tendency of the solvent may be reduced in consequence of the reduction in the number of free molecules of solvent; thus, a negative deviation from Raoult's law is introduced. On the other hand, the escaping-tendency of the solvent, in a solution of nonvolatile solute, may be increased because the cohesive forces between molecules of solvent are reduced by the solute; this results in a positive deviation from Raoult's law. Chapter 17 considers deviations from ideality in more detail.

While few solutions exhibit ideal behavior over a wide range of concentration, most solutions behave ideally at least in high dilution, where deviations from Raoult's law are negligible.

Comparison of Colligative Properties—In view of the established interrelationships of the colligative properties of ideal solutions or very dilute real solutions, it is possible to predict, by calculation, the magnitude of all these properties of such solutions if the concentration of the nonelectrolyte solute is given. Also, if the magnitude of one of the properties, eg, the freezing point, is known for a solution of unspecified concentration, it is possible to calculate the vapor pressure, boiling point and osmotic pressure, provided the solution is ideal or sufficiently dilute to show negligible deviation from ideality. To what upper limit of concentration a nonideal solution remains "sufficiently dilute" to show ideal behavior is difficult to specify. The answer depends at least in part on the degree of agreement expected between experimental and theoretical values. Certainly, a 1-molal concentration is much too concentrated for a nonideal solution to show conformance with ideal behavior and even in 0.1-molal concentration, deviations are significant and for some purposes may be excessive.

In dealing with colligative properties of solutions which do not behave ideally, caution should be exercised in attempting to predict the magnitude of other colligative properties from one that has been determined experimentally. Earlier, an equation was derived for calculating the vapor pressure of a solution from its osmotic pressure, or *vice versa*, this equation being valid even with solutions showing substantial departures from ideal behavior. The equation is limited, however, to a comparison of these colligative properties at the same temperature. The degree of deviation from ideal behavior for one colligative property will be exactly the same for another only when the temperature is the same for both. It does not follow that the degree of deviation of the colligative properties of a given nonideal solution will be the same for all the properties since at least two of these (freezing point and boiling point) must be determined at quite different temperatures. While in dilute solutions the intermolecular (and/or interionic) forces and interactions *may* change little over the temperature interval between freezing and boiling, in concentrated solutions the change *may* be marked. In the absence of adequate knowledge about the forces and interactions involved, only by experiment can one establish the magnitude of the colligative properties of other than very dilute nonideal solutions. It is important to keep this in mind in estimating the osmotic pressure of a nonideal solution at body temperature from a freezing point determined some 37° lower. While in many cases—possibly the majority of them—such an estimate is warranted by virtue of essential constancy of the various forces and interactions over a wide range of temperature, this is not always the case and the estimate may be significantly inaccurate.

Colligative Properties of Electrolyte Solutions [See Chapter 17]—Earlier in this chapter attention was directed to the increased osmotic pressure observed in solutions of electrolytes, the enhanced effect being attributed to the presence of ions, each of which acts, in general, in the same way as a molecule in developing osmotic pressure. Similar magnification of vapor-pressure lowering, boiling-point elevation and freezing-point depression occurs in solutions of electrolytes. Thus, at a given constant temperature the abnormal effect of an electrolyte on osmotic pressure is paralleled by abnormal lowering of vapor pressure; the other colligative properties are, subject to variation of effect with temperature, comparably intensified. In general the magnitude of each colligative property is proportional to the total number of particles (molecules and/or ions) in solution.

While in *very* dilute solutions the osmotic pressure, vapor-pressure lowering, boiling-point increase and freezing-point depression of solutions of electrolytes would approach values 2, 3, 4, etc times greater for NaCl, Na_2SO_4 and Na_3PO_4 than in solutions of the same molality of a nonelectrolyte, two other effects are observed as the concentration of electrolyte is increased. The first effect results in *less* than 2-, 3- or 4-fold intensification of a colligative property. This reduction is ascribed to interionic attraction between the positively and negatively charged ions, in consequence of which the ions are not completely dissociated from each other and do not exert their full effect in lowering vapor pressure, etc. This deviation generally increases with increasing concentration of electrolyte. The second effect intensifies the colligative properties and is attributed to the attraction of ions for solvent molecules, which holds the solvent in solution and reduces its escaping-tendency, with consequent enhancement of the vapor-pressure lowering. Solvation also may reduce interionic attraction and thereby further lower the vapor pressure. These factors (and possibly others) combine to effect a progressive reduction in the molal values of colligative properties as the concentration of electrolyte is increased to 0.5 to 1 molal, beyond which the molal quantities either increase, sometimes quite abruptly, or remain almost constant.

Activity and Activity Coefficient

Various mathematical expressions are employed to relate properties of chemical systems (equilibrium constants, colligative properties, pH, etc) to the stoichiometric concentration of one or more molecular, atomic or ionic species. In deriving such expressions it either is stated or implied that they are valid only so long as intermolecular, interatomic and/or interionic forces may be ignored or remain constant, under which restriction the system may be expected to behave ideally. But intermolecular, interatomic and/or interionic forces do exist, and not only do they change as a result of chemical reaction but also with variation in the concentration or pressure of the molecules, atoms or ions under observation. In consequence, mathematical expressions involving stoichiometric concentrations or pressures generally have limited applicability. The conventional concentration terms, while providing a count of molecules, atoms or ions per unit volume, afford no indication of the physical or chemical activity of the species measured, and it is this activity which determines the physical and chemical properties of the system.

In recognition of this, GN Lewis introduced both the quantitative concept and methods for evaluation of activity as a true measure of the physical or chemical activity of molecular, atomic or ionic species, whether in the state of gas, liquid or solid, or whether present as a single species or in a mixture. Activity may be considered loosely as a corrected concentration or pressure which takes into account not only the stoichiometric concentration or pressure but also any intermolecular attractions, repulsions or interactions between solute and solvent in solution, association and ionization. Thus, activity measures the net effectiveness of a chemical species. Because only relative values of activity may be determined, a *standard state* must be chosen for quantitative comparisons to be made. Indeed, because activity measurements are needed for many different types of systems, several standard states must be selected. Since this discussion is concerned mainly with solutions, the standard state for the solvent is pure solvent, while for the solute it is a hypothetical solution with free energy corresponding to unit molality under conditions of ideal behavior of the solution. The relationship of activity to concentration is measured in terms of an activity coefficient which is discussed in Chapter 17.

Practical Applications of Colligative Properties

One of the most important pharmaceutical applications of colligative properties is in the preparation of isotonic intravenous and isotonic lacrimal solutions, the details of which are discussed in Chapter 79.

Other applications of the colligative properties are found in experimental physiology. One such application is in the immersion of tissues in salt solutions, which are isotonic with the fluids of the tissue, in order to prevent changes or injuries that may arise from osmosis.

The colligative properties of solutions also may be used in determining the molecular weight of solutes or, in the case of electrolytes, the extent of ionization. The method of determining molecular weight depends on the fact that each of the colligative properties is altered by a constant value when a definite number of molecules of solute is added to a solvent [See Chapter 17]. For example, in dilute solutions the freezing point of water is lowered at the rate of 1.855° for each gram-molecular weight of a nonelectrolyte dissolved in 1000 g of water.*

The constant 1.855° is commonly called the *molal freezing-point depression* of water. To find the molecular weight of a nonelectrolyte, therefore, all that is necessary is to determine the freezing point of a dilute aqueous solution of known concentration of the nonelectrolyte and, by proportion, to calculate the quantity that would produce, theoretically, a depression of 1.855° when 1000 g of water is used as the solvent. If the substance is insoluble in water, it may be dissolved in another solvent, in which case, however, the freezing-point depression of a solution corresponding to a gram-molecular weight of the solute in 1000 g of solvent will be some value other than 1.855°. In the case of benzene, for example, this value is 5.12°; carbon tetrachloride, 2.98°; phenol, 7.27° and camphor, about 40° (see *Freezing-Point Depression*, in this chapter).

The boiling-point elevation may be used similarly for determining molecular weights. The boiling point of water is raised at the rate of 0.52° for each gram-molecular weight of solute dissolved in 1000 g of water,* the corresponding values for benzene, carbon tetrachloride and phenol being 2.57°, 4.88° and 3.60°, respectively. The observation of vapor-pressure lowering and osmotic pressure likewise may be used to calculate molecular weights.

To determine the extent to which an electrolyte is ionized, it is necessary to know its molecular weight, as determined by some other method, and then to measure one of the four colligative properties. The deviation of the results from similar values for nonelectrolytes then is used in calculating the extent of ionization.

References

1. Martin AN *et al: Physical Pharmacy*, Lea & Febiger, Philadelphia, 1983.
2. Kramer SF, Flynn GL: *J Pharm Sci 61:* 1896, 1972.
3. Campbell AN, Campbell AJR: *J Am Chem Soc 59:* 2481, 1937.
4. Higuchi T, in Lyman R: *Pharmaceutical Compounding and Dispensing*, Lippincott, Philadelphia, 159, 1949.
5. Yalkowsky, SH: *Techniques of Solubilization of Drugs*, Marcel Dekker, New York, 91, 1981.
6. Martin A *et al: J Pharm Sci, 71:* 849, 1982.
7. Setnikar I, Temelcou O: *JAPhA Sci Ed 48:* 628, 1959.

* These constants apply only to solutions that are considerably more dilute than 1 molal; a substantial deviation would be observed if a 1-molal solution were to be used.

CHAPTER 17

Ionic Solutions and Electrolytic Equilibria

Paul J Niebergall, PhD

Professor of Pharmaceutical Sciences and Director
Pharmaceutical Development Center
Medical University of South Carolina
Charleston, SC 29425

Electrolytes

In the preceding chapter, attention was directed to the colligative properties of nonelectrolytes, or substances whose aqueous solutions do not conduct electricity. Substances whose aqueous solutions conduct electricity are known as electrolytes, and are typified by inorganic acids, bases and salts. In addition to the property of electrical conductivity, solutions of electrolytes exhibit anomalous colligative properties.

Colligative Properties

In general, for nonelectrolytes, a given colligative property of two equimolal solutions will be identical. This generalization, however, cannot be made for solutions of electrolytes.

Van't Hoff pointed out that the osmotic pressure of a solution of an electrolyte is considerably greater than the osmotic pressure of a solution of a nonelectrolyte of the same molal concentration. This anomaly remained unexplained until 1887 when Arrhenius proposed a hypothesis which forms the basis for our modern theories of electrolyte solutions.

This theory postulated that when electrolytes are dissolved in water they split up into charged particles known as ions. Each of these ions carries one or more electrical charges, with the total charge on the positive ions (cations) being equal to the total charge on the negative ions (anions). Thus, although a solution may contain charged particles, it remains neutral. The increased osmotic pressure of such solutions is due to the increased number of particles formed in the process of ionization. For example, sodium chloride is assumed to dissociate as

$$Na^+Cl^- \xrightarrow{H_2O} Na^+ + Cl^-$$

It is evident that each molecule of sodium chloride which is dissociated produces two ions and, if dissociation is complete, there will he twice as many particles as would be the case if it were not dissociated at all. Furthermore, if each ion has the same effect on osmotic pressure as a molecule, it might be expected that the osmotic pressure of the solution would be twice that of a solution containing the same molal concentration of a nonionizing solute.

Osmotic-pressure data indicate that, in very dilute solutions of salts which yield two ions, the pressure is very nearly double that of solutions of equimolal concentrations of nonelectrolytes. Similar magnification of vapor-pressure lowering, hoiling-point elevation and freezing-point depression occurs in dilute solutions of electrolytes.

Van't Hoff defined a factor i as the ratio of the colligative effect produced hy a concentration m of electrolyte, divided hy the effect observed for the same concentration of nonelectrolyte, or

$$i = \frac{\pi}{(\pi)_0} = \frac{\Delta P}{(\Delta P)_0} = \frac{\Delta T_b}{(\Delta T_b)_0} = \frac{\Delta T_f}{(\Delta T_f)_0} \qquad (1)$$

in which π, ΔP, ΔT_b and ΔT_f refer to the osmotic pressure, vapor-pressure lowering, boiling-point elevation and freezing-point depression, respectively, of the electrolyte. The terms $(\pi)_0$, etc refer to the nonelectrolyte of the same concentration. In general, with strong electrolytes (those assumed to be 100% ionized), the van't Hoff factor is equal to the number of ions produced when the electrolyte goes into solution (2 for NaCl and $MgSO_4$, 3 for $CaCl_2$ and Na_2SO_4, 4 for $FeCl_3$ and Na_3PO_4, etc).

In *very* dilute solutions the osmotic pressure, vapor-pressure lowering, boiling-point elevation and freezing-point depression of solutions of electrolytes approach values 2, 3, 4 or more times greater (depending on the type of strong electrolyte) than in solutions of the same molality of nonelectrolyte, thus confirming the hypothesis that an ion has the same primary effect as a molecule on colligative properties. It bears repeating, however, that two other effects are observed as the concentration of electrolyte is increased.

The first effect results in less than 2-, 3- or 4-fold intensification of a colligative property. This reduction is ascribed to interionic attraction between the positively and negatively charged ions, in consequence of which the ions are not dissociated completely from each other and do not exert their full effect on vapor pressure and other colligative properties. This deviation generally increases with increasing concentration of electrolyte.

The second effect intensifies the colligative properties and is attributed to the attraction of ions for solvent molecules (called solvation, or, if water is the solvent, hydration), which holds the solvent in solution and reduces its escaping tendency, with a consequent enhancement of the vapor-pressure lowering. Solvation also reduces interionic attraction and, thereby, further lowers the vapor pressure.

Conductivity

The ahility of metals to conduct an electric current results from the mobility of electrons in the metals. This type of conductivity is called *metallic* conductance. On the other hand, various chemical compounds—notahly acids, bases and salts—conduct electricity by virtue of ions present or formed, rather than by electrons. This is called *electrolytic* conductance, and the conducting compounds are electrolytes. While the fact that certain electrolytes conduct electricity in the molten state is important, their behavior when dissolved in a solvent, particularly in water, is of greater concern in pharmaceutical science.

The electrical conductivity (or conductance) of a solution of an electrolyte is merely the reciprocal of the resistance of the solution. Hence, to measure conductivity is actually to

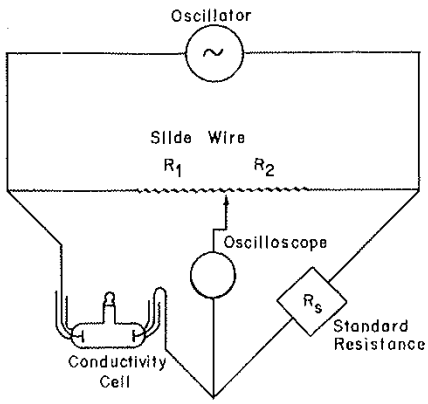


Fig 17-1. Alternating current Wheatstone bridge for measuring conductivity.

measure electrical resistance, commonly with a Wheatstone-bridge apparatus, and then to *calculate* the conductivity. Fig 17-1 is a representation of the component parts of the apparatus.

The solution to be measured is placed in a glass or quartz cell having two inert electrodes, commonly made of platinum or gold and coated with spongy platinum to absorb gases, across which passes an alternating current generated by an oscillator at a frequency of about 1000 Hz. The reason for using alternating current is to reverse the electrolysis that occurs during flow of current and which would cause polarization of the electrodes and lead to abnormal results. The size of the electrodes and their distance apart may be varied to reduce very high resistance and increase very low resistance in order to increase the accuracy and precision of measurement. Thus, solutions of high conductance (low resistance) are measured in cells having small electrodes relatively far apart, while solutions of low conductance (high resistance) are measured in cells with large electrodes placed close to each other.

Electrolytic resistance, like metallic resistance, varies directly with the length of the conducting medium and inversely with its cross-sectional area. The known resistance required for the circuit is provided by a resistance box containing calibrated coils. Balancing of the bridge may be achieved by sliding a contact over a wire of uniform resistance until no (or minimum) current flows through the circuit, as detected either visually with a cathode-ray oscilloscope or audibly with earphones.

The resistance, in ohms, is calculated by the simple procedure used in the Wheatstone-bridge method. The reciprocal of the resistance is the conductivity, the units of which are *reciprocal ohms* (also called *mhos*). As the numerical value of the conductivity will vary with the dimensions of the conductance cell, the value must be calculated in terms of *specific conductance*, L, which is the conductance in a cell having electrodes of 1-sq cm cross-sectional area and 1 cm apart. If the dimensions of the cell used in the experiment were known, it would be possible to calculate the specific conductance, but this information actually is not required, because it is possible—and much more convenient—to calibrate a cell by measuring in it the conductivity of a standard solution of known specific conductance and then calculating a "cell constant." Since this constant is a function only of the dimensions of the cell, it can be used to convert all measurements in that cell to specific conductivity. Solutions of known concentration of pure potassium chloride are used as standard solutions for this purpose.

Equivalent Conductance—In studying the variation of conductance of electrolytes with dilution it is essential to make allowance for the degree of dilution in order that the comparison of conductances may be made for identical amounts of solute. This may be achieved by expressing conductance measurements in terms of *equivalent conductance*, Λ, which is obtained by multiplying the specific conductance, L, by the volume in milliliters, V_e, of solution containing 1 g-eq of solute. Thus

$$\Lambda = LV_e = \frac{1000L}{C} \qquad (2)$$

where C is the concentration of electrolyte in the solution in g-Eq/L, ie, the normality of the solution. For example, the equivalent conductance of 0.01 N potassium chloride solution, which has a specific conductance of 0.001413 mho/cm may be calculated in either of the following ways

$$\Lambda = 0.001413 \times 100{,}000 = 141.3 \text{ mho cm}^2\text{/eq}$$

or

$$\Lambda = \frac{1000 \times 0.001413}{0.01} = 141.3$$

Strong and Weak Electrolytes—It is customary to classify electrolytes broadly as *strong electrolytes* and *weak electrolytes*. The former category includes solutions of strong acids, strong bases and most salts; the latter includes weak acids and bases, primarily organic acids, amines and a few salts. The usual criterion for distinguishing between strong and weak electrolytes is the extent of ionization. An electrolyte existing entirely or very largely as ions is considered a strong electrolyte, while one that is a mixture of a substantial proportion of molecular species along with ions derived therefrom is a weak electrolyte. For the purposes of this discussion, classification of electrolytes as strong or weak will be on the basis of certain conductance characteristics exhibited in aqueous solution.

The equivalent conductances of a number of electrolytes, at different concentrations, are given in Table I and for certain of these electrolytes again in Fig 17-2, where the

Table I—Equivalent Conductances at 25°

g-Eq/L	HCl	HAc	NaCl	KCl	NaI	KI	NaAc
Inf dil	426.1	390.6[a]	126.5	149.9	126.9	150.3	91.0
0.0005	422.7	67.7	124.5	147.8	125.4	...	89.2
0.001	421.4	49.2	123.7	146.9	124.3	...	88.5
0.005	415.8	22.9	120.6	143.5	121.3	144.4	85.7
0.01	412.0	16.3	118.5	141.3	119.2	142.2	83.8
0.02	407.2	11.6	115.8	138.3	116.7	139.5	81.2
0.05	399.1	7.4	111.1	133.4	112.8	135.0	76.9
0.1	391.3	5.2	106.7	129.0	108.8	131.1	72.8

[a] The equivalent conductance at infinite dilution for acetic acid, a weak electrolyte, is obtained by adding the equivalent conductances of hydrochloric acid and sodium acetate and subtracting that of sodium chloride (see text for explanation).

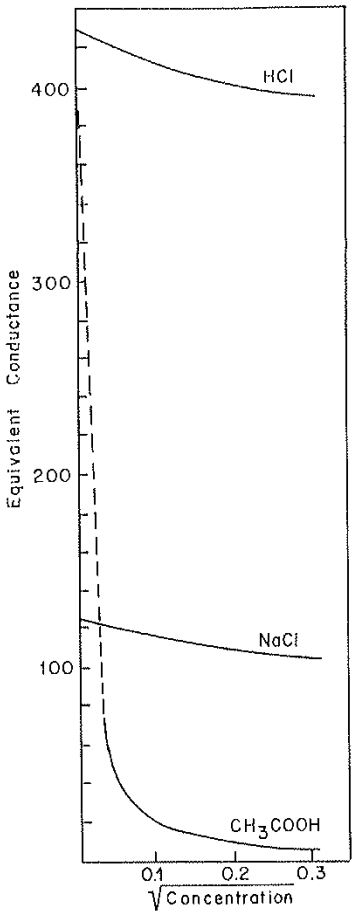


Fig 17-2. Variation of equivalent conductance with square root of concentration.

equivalent conductance is plotted against the square root of concentration. By plotting the data in this manner a linear relationship is observed for strong electrolytes, while a steeply rising curve is noted for weak electrolytes; this difference is a characteristic which distinguishes strong and weak electrolytes. The interpretation of the steep rise in the equivalent conductance of weak electrolytes is that the degree of ionization increases with dilution, becoming complete at infinite dilution. Interionic interference effects generally have a minor role in the conductivity of weak electrolytes. With strong electrolytes, which usually are completely ionized, the increase in equivalent conductance results not from increased ionization but rather from diminished ionic interference as the solution is diluted, in consequence of which ions have greater freedom of mobility, ie, increased conductance.

The value of the equivalent conductance extrapolated to infinite dilution (zero concentration), designated by the symbol Λ_0, has a special significance. It represents the equivalent conductance of the completely ionized electrolyte when the ions are so far apart that there is no interference with their migration due to interionic interactions. It has been shown, by Kohlrausch, that the equivalent conductance of an electrolyte at infinite dilution is the sum of the equivalent conductances of its component ions at infinite dilution, expressed symbolically as

Table II—Equivalent Ionic Conductivities at Infinite Dilution, at 25°

Cations	l_0	Anions	l_0
H^+	349.8	OH^-	198.0
Li^+	38.7	Cl^-	76.3
Na^+	50.1	Br^-	78.4
K^+	73.5	I^-	76.8
NH_4^+	61.9	Ac^-	40.9
$\frac{1}{2}Ca^{2+}$	59.5	$\frac{1}{2}SO_4^{2-}$	79.8
$\frac{1}{2}Mg^{2+}$	53.0		

$$\Lambda_0 = l_0(\text{cation}) + l_0(\text{anion}) \qquad (3)$$

The significance of Kohlrausch's law is that each ion, at infinite dilution, has a characteristic value of conductance that is independent of the conductance of the oppositely charged ion with which it is associated. Thus, if the equivalent conductances of various ions are known, the conductance of any electrolyte may be calculated simply by adding the appropriate ionic conductances. Since the fraction of current carried by cations (*transference number* of the cations) and by anions (transference number of anions) in an electrolyte may be determined readily by experiment, ionic conductances are known. Table II gives the equivalent ionic conductances at infinite dilution of some cations and anions. It is not necessary to have this information in order to calculate the equivalent conductance of an electrolyte, for Kohlrausch's law permits the latter to be calculated by adding and subtracting values of Λ_0 for appropriate electrolytes. For example, the value of Λ_0 for acetic acid may be calculated as

$$\Lambda_0(CH_3COOH) = \Lambda_0(HCl) + \Lambda_0(CH_3COONa) - \Lambda_0(NaCl)$$

which is equivalent to

$$l_0(H^+) + l_0(CH_3COO^-) = l_0(H^+) + l_0(Cl^-) + l_0(Na^+) + l_0(CH_3COO^-) - l_0(Na^+) - l_0(Cl^-)$$

This method is especially useful for calculating Λ_0 for weak electrolytes such as acetic acid. As is evident from Fig 17-2, the Λ_0 value for acetic acid cannot be determined accurately by extrapolation because of the steep rise of conductance in dilute solutions. For strong electrolytes, on the other hand, the extrapolation can be made very accurately. Thus, in the example above, the values of Λ_0 for HCl, CH_3COONa and NaCl are determined easily by extrapolation since the substances are strong electrolytes. Substitution of these extrapolated values, as given in Table II, yields a value of 390.6 for the value of Λ_0 for CH_3COOH.

Ionization of Weak Electrolytes—When Arrhenius introduced his theory of ionization he proposed that the degree of ionization, α, of an electrolyte is measured by the ratio

$$\alpha = \Lambda/\Lambda_0 \qquad (4)$$

where Λ is the equivalent conductance of the electrolyte at any specified concentration of solution and Λ_0 is the equivalent conductance at infinite dilution. As strong electrolytes were not then recognized as being 100% ionized, and interionic interference effects had not been evaluated, he believed the equation to be applicable to both strong and weak electrolytes. Since it now is known that the apparent variation of ionization of strong electrolytes arises from a change in the mobility of ions at different concentrations, rather than from varying ionization, the equation is not applicable to strong electrolytes. It does provide, however, a generally acceptable approximation of the degree of ionization of weak

electrolytes, for which deviations resulting from neglect of activity coefficients and of some change of ionic mobilities with concentration are, for most purposes, negligible. The following example illustrates the use of the equation to calculate the degree of ionization of a typical weak electrolyte.

Example—Calculate the degree of ionization of 1×10^{-3} *N* acetic acid, the equivalent conductance of which is 48.15 mho cm²/Eq. The equivalent conductance at infinite dilution is 390.6 mho cm²/Eq.

$$\alpha = \frac{48.15}{390.6} = 0.12$$

$$\% \text{ ionization} = 100\alpha = 12\%$$

The degree of dissociation also can he calculated using the van't Hoff factor, i, and the following equation

$$\alpha = \frac{i - 1}{\upsilon - 1} \tag{5}$$

where υ is the number of ions into which the electrolyte dissociates.

Example—A 1.0×10^{-3} *N* solution of acetic acid has a van't Hoff factor equal to 1.12. Calculate the degree of dissociation of the acid at this concentration.

$$\alpha = \frac{i - 1}{\upsilon - 1} = \frac{1.12 - 1}{2 - 1} = 0.12$$

This result agrees with that obtained using equivalent condoctance and Eq 4.

Modern Theories

The Arrhenius theory explains why solutions of electrolytes conduct electricity, why they exhibit enhanced colligative properties and essentially is satisfactory for solutions of weak electrolytes. Several deficiencies, however, do exist when it is applied to solutions of strong electrolytes. It does not explain the failure of strong electrolytes to follow the law of mass action as applied to ionization; discrepancies exist between the degree of ionization calculated from the van't Hoff factor and the conductivity ratio for strong electrolyte solutions having concentrations greater than about 0.5 *M*.

These deficiencies can be explained hy the following observations

1. In the molten state, strong electrolytes are excellent conductors of electricity. This suggests that these materials are already ionized in the crystalline state. Further support for this is given by X-ray studies of crystals, which indicate that the units comprising the basic lattice structure of strong electrolytes are ions.
2. Arrhenius neglected the fact that ions in solution, being oppositely charged, tend to associate through electrostatic attraction. In solutions of weak electrolytes, the number of ions is not large and it is not surprising that electrostatic attractions do not cause appreciable deviations from theory. In dilute solutions, in which strong electrolytes are assumed to be 100% ionized, the number of ions is large, and interionic attractions become major factors in determining the chemical properties of these solutions. These effects should, and do, become more pronounced as the concentration of electrolyte or the valence of the ions is increased.

It is not surprising, therefore, that the Arrhenius theory of partial ionization involving the law of mass action and neglecting ionic charge does not hold for solutions of strong electrolytes. Neutral molecules of strong electrolytes, if they do exist in solution, must arise from interionic attraction rather than from incomplete ionization.

Activity and Activity Coefficients—Due to increased electrostatic attractions as a solution becomes more concentrated, the concentration of an ion becomes less efficient as a measure of its net effectiveness. A more efficient measure of the physical or chemical effectiveness of an ion is known as its *activity*, which is a measure of an ion's concentration related to its concentration at a universally adopted reference-standard state. The relationship between the activity and the concentration of an ion can be expressed as

$$a = m\gamma \tag{6}$$

where m is the molal concentration, γ is the activity coefficient and a is the activity. The activity also can be expressed in terms of molar concentration, c, as

$$a = fc \tag{7}$$

where f is the activity coefficient on a molar scale. In dilute solutions (below 0.01 *M*) the two activity coefficients are identical, for all practical purposes.

The activity coefficient may be determined in various ways, such as measuring colligative properties, electromotive force, solubility, distribution coefficients, etc. For a strong electrolyte, the mean ionic activity coefficient, $\gamma_\pm$, or $f_\pm$, provides a measure of the deviation of the electrolyte from ideal behavior. The mean ionic activity coefficients on a molal basis for several strong electrolytes are given in Tahle II. It is characteristic of the electrolytes that the coefficients at first decrease with increasing concentration, pass through a minimum and finally increase with increasing concentration of electrolyte.

Ionic Strength—Ionic strength is a measure of the intensity of the electrical field in a solution and may be expressed as

$$\mu = \tfrac{1}{2}\Sigma c_i z_i^2 \tag{8}$$

where z_i is the valence of ion i. The mean ionic activity coefficient is a function of ionic strength as are such diverse phenomena as solubilities of sparingly soluble substances, rates of ionic reactions, effects of salts on pH of buffers, electrophoresis of proteins, etc.

The greater effectiveness of ions of higher charge type on a specific property, compared with the effectiveness of the same number of singly charged ions, generally coincides with the ionic strength calculated by Eq 8. The variation of ionic strength with the valence (charge) of the ions comprising a strong electrolyte should he noted.

For univalent cations and univalent anions (called uniunivalent or 1-1) electrolytes, the ionic strength is identical with molarity. For bivalent cation and univalent anion (biunivalent or 2-1) electrolytes, or univalent cation and bivalent anion (unibivalent or 1-2) electrolytes, the ionic strength is three times the molarity. For bivalent cation and bivalent anion (bibivalent or 2-2) electrolytes, the ionic strength is four times the molarity. These relationships are evident from the following example.

Example—Calculate the ionic strength of 0.1 *M* solutions of NaCl, Na_2SO_4, $MgCl_2$ and $MgSO_4$, respectively. For

$$\text{NaCl} \quad \mu = \tfrac{1}{2}(0.1 \times 1^2 + 0.1 \times 1^2) = 0.1$$

$$\text{Na}_2\text{SO}_4 \quad \mu = \tfrac{1}{2}(0.2 \times 1^2 + 0.1 \times 2^2) = 0.3$$

$$\text{MgCl}_2 \quad \mu = \tfrac{1}{2}(0.1 \times 2^2 + 0.2 \times 1^2) = 0.3$$

$$\text{MgSO}_4 \quad \mu = \tfrac{1}{2}(0.1 \times 2^2 + 0.1 \times 2^2) = 0.4$$

The ionic strength of a solution containing more than one electrolyte is the sum of the ionic strengths of the individual salts comprising the solution. For example, the ionic strength of a solution containing NaCl, Na_2SO_4, $MgCl_2$ and $MgSO_4$, each at a concentration of 0.1 *M*, is 1.1.

Debye-Hückel Theory—The Dehye-Hückel equations which are applicable only to very dilute solutions (about 0.02 μ), may be extended to somewhat more concentrated solutions (about 0.1 μ) in the simplified form

$$\log f_i = \frac{-0.51\, z_1^2 \sqrt{\mu}}{1 + \sqrt{\mu}} \tag{9}$$

Table III—Values of Some Salting-Out Constants for Various Barbiturates at 25°

Barbiturate	KCl	KBr	NaCl	NaBr
Amobarbital	0.168	0.095	0.212	0.143
Aprobarbital	0.136	0.062	0.184	0.120
Barbital	0.092	0.042	0.136	0.088
Phenobarbital	0.092	0.034	0.132	0.078
Vinbarbital	0.125	0.036	0.143	0.096

The mean ionic activity coefficient for aqueous solutions of electrolytes at 25° can be expressed as

$$\log f_{\pm} = \frac{-0.51\, z_+ z_- \sqrt{\mu}}{1 + \sqrt{\mu}} \qquad (10)$$

in which z_+ is the valence of the cation and z_- is the valence of the anion. When the ionic strength of the solution becomes high (approximately 0.3 to 0.5), these equations become inadequate and a linear term in μ is added. This is illustrated for the mean ionic activity coefficient

$$\log f_{\pm} = \frac{-0.51\, z_+ z_- \sqrt{\mu}}{1 + \sqrt{\mu}} + K_s \mu \qquad (11)$$

in which K_s is a "salting-out" constant chosen empirically for each salt. This equation is valid for solutions with ionic strength up to approximately 1.

Salting-Out Effect—The aqueous solubility of a slightly soluble organic substance generally is affected markedly by the addition of an electrolyte. This effect particularly is noticeable when the electrolyte concentration reaches 0.5 *M* or higher. If the aqueous solution of the organic substance has a dielectric constant lower than that of pure water, its solubility is decreased and the substance is "salted-out." The use of high concentrations of electrolytes, such as ammonium sulfate or sodium sulfate, for the separation of proteins by differential precipitation is perhaps the most striking example of this effect. The aqueous solutions of a few substances such as hydrocyanic acid, glycine and cystine have a higher dielectric constant than that of pure water, and these substances are "salted-in." These phenomena can be expressed empirically as

$$\log S = \log S_0 \pm K_s m \qquad (12)$$

in which S_0 represents the solubility of the organic substance in pure water and S is the solubility in the electrolyte solution. The slope of the straight line obtained by plotting log S versus m is positive for "salting-in" and negative for salting-out. In terms of ionic strength this equation becomes

$$\log S = \log S_0 \pm K_s' \mu \qquad (13)$$

where $K_s' = K_s$ for univalent salts, $K_s' = K_s/3$ for unibivalent salts and $K_s' = K_s/4$ for bivalent salts. The salting-out constant depends on the temperature as well as the nature of both the organic substance and the electrolyte. The effect of the electrolyte and the organic substance can be seen in Table III. In all instances, if the anion is constant, the sodium cation has a greater salting-out effect than the potassium cation, probably due to the higher charge density of the former. Although the reasoning is less clear, it appears that for a constant cation, chloride anion has a greater effect than bromide anion upon the salting-out phenomenon.

Acids and Bases

Arrhenius defined an acid as a substance that yields hydrogen ions in aqueous solution and a base as a substance that yields hydroxyl ions in aqueous solution. Except for the fact that hydrogen ions neutralize hydroxyl ions to form water, no complementary relationship between acids and bases (such as that between oxidants and reductants, for example) is evident in Arrhenius' definitions for these substances; rather, their oppositeness of character is emphasized. Moreover, no account is taken of the behavior of acids and bases in nonaqueous solvents. Also, while acidity is associated with so elementary a particle as the proton (hydrogen ion), basicity is attributed to so relatively complex an association of atoms as the hydroxyl ion. It would seem that a simpler concept of a base could be devised.

Proton Concept—In pondering the objections to Arrhenius' definitions, Brønsted and Bjerrum in Denmark and Lowry in England developed, and in 1923 announced, a more satisfactory, and more general, theory of acids and bases. According to this theory an acid is a substance capable of yielding a proton (hydrogen ion), while a base is a substance capable of accepting a proton. This complementary relationship may be expressed by the general equation

$$\underset{\text{acid}}{A} \rightleftharpoons H^+ + \underset{\text{base}}{B}$$

The pair of substances thus related through mutual ability to gain or lose a proton is called a *conjugate acid–base pair.* Specific examples of such pairs are

Acid		Base
HCl	$\rightleftharpoons$	$H^+ + Cl^-$
CH_3COOH	$\rightleftharpoons$	$H^+ + CH_3COO^-$
NH_4^+	$\rightleftharpoons$	$H^+ + NH_3$
HCO_3^-	$\rightleftharpoons$	$H^+ + CO_3^{2-}$
$H_2PO_4^-$	$\rightleftharpoons$	$H^+ + HPO_4^{2-}$
H_2O	$\rightleftharpoons$	$H^+ + OH^-$
H_3O^+	$\rightleftharpoons$	$H^+ + H_2O$
$Al(H_2O)_6^{3+}$	$\rightleftharpoons$	$H^+ + Al(H_2O)_5OH^{2+}$

It is apparent that not only molecules but also cations and anions may function as acids or bases.

The complementary nature of the acid–base pairs listed is reminiscent of the complementary relationship of pairs of oxidants and reductants where, however, the ability to gain or lose one or more electrons—rather than protons—is the distinguishing characteristic.

Oxidant		Reductant
$Fe^{3+} + e^-$	$\rightleftharpoons$	Fe^{2+}
$Na^+ + e^-$	$\rightleftharpoons$	Na
$\frac{1}{2}I_2 + e^-$	$\rightleftharpoons$	I^-

However, these examples of acid–base pairs and oxidant–reductant pairs represent reactions that are possible in principle only. Ordinarily acids will not release free protons any more than reductants will release free electrons. That is, protons and electrons, respectively, can be *transferred* only from one substance (an ion, atom or molecule) to another. Thus, it is a fundamental fact of chemistry that oxidation of one substance will occur only if reduction of another substance occurs simultaneously. Stated in another way, electrons will be released from the reductant (oxidation) only if an oxidant capable of accepting electrons (reduction) is present. For this reason oxidation–reduction reactions must involve two conjugate oxidant–reductant pairs of substances

$$\text{oxidant}_1 + \text{reductant}_2 \rightleftharpoons \text{reductant}_1 + \text{oxidant}_2$$

where Subscript 1 represents one conjugate oxidant–reductant pair and Subscript 2 represents the other.

Similarly, an acid will not release a proton unless a base capable of accepting it is present simultaneously. This means that any actual manifestation of acid–base behavior must involve interaction between two sets of conjugate acid–base pairs, represented as

$$\underset{\text{acid}_1}{A_1} + \underset{\text{base}_2}{B_2} \rightleftharpoons \underset{\text{base}_1}{B_1} + \underset{\text{acid}_2}{A_2}$$

In such a reaction, which is called *protolysis* or a *protolytic reaction*, A_1 and B_1 constitute one conjugate acid–base pair and A_2 and B_2 the other; the proton given up by A_1 (which thereby becomes B_1) is transferred to B_2 (which becomes A_2).

When an acid, such as hydrochloric, is dissolved in water, a protolytic reaction occurs.

$$\underset{\text{acid}_1}{HCl} + \underset{\text{base}_2}{H_2O} \rightleftharpoons \underset{\text{base}_1}{Cl^-} + \underset{\text{acid}_2}{H_3O^+}$$

The ionic species H_3O^+, called *hydronium* or *oxonium* ion, always is formed when an acid is dissolved in water. Often, for purposes of convenience, this is written simply as H^+ and is called hydrogen ion, although the "bare" ion practically is nonexistent in solution.

When a base, eg, ammonia, is dissolved in water the reaction of protolysis is

$$\underset{\text{base}_1}{NH_3} + \underset{\text{acid}_2}{H_2O} \rightleftharpoons \underset{\text{acid}_1}{NH_4^+} + \underset{\text{base}_2}{OH^-}$$

The proton theory of acid–base function makes the concept of hydrolysis superfluous. When, for example, sodium acetate is dissolved in water, this acid–base interaction occurs

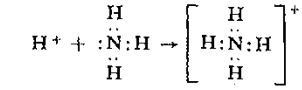
$$\underset{\text{base}_1}{CH_3COO^-} + \underset{\text{acid}_2}{H_2O} \rightleftharpoons \underset{\text{acid}_1}{CH_3COOH} + \underset{\text{base}_2}{OH^-}$$

In an aqueous solution of ammonium chloride the reaction is

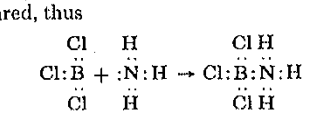
$$\underset{\text{acid}_1}{NH_4^+} + \underset{\text{base}_2}{H_2O} \rightleftharpoons \underset{\text{base}_1}{NH_3} + \underset{\text{acid}_2}{H_3O^+}$$

Transfer of protons (protolysis) is not limited to dissimilar conjugate acid–base pairs. In the preceding examples H_2O sometimes behaves as an acid and at other times as a base. Such an amphoteric substance is called, in Brønsted's terminology, an *amphiprotic substance*.

Electron-Pair Concept—While the proton concept of acids and bases provides a more general definition for these substances, it does not indicate the basic reason for proton transfer, nor does it explain how such substances as sulfur trioxide, boron trichloride, stannic chloride or carbon dioxide—none of which is capable of donating a proton—can behave as acids. Both deficiencies of the proton theory are avoided in the more inclusive definition of acids and bases proposed by Lewis in 1923. In 1916 he proposed that sharing of a pair of electrons by two atoms established a bond (covalent) between the atoms; therefore, an acid is a substance capable of sharing a pair of electrons made available by another substance called a base, thereby forming a coordinate covalent bond. The base is the substance that donates a share in its electron pair to the acid. The following equation illustrates how Lewis' definitions explain the transfer of a proton (hydrogen ion) to ammonia to form ammonium ion.



The reaction of boron trichloride, which according to the Lewis theory is an acid, with ammonia is similar, for the boron lacks an electron pair if it is to attain a stable octet configuration, while ammonia has a pair of electrons which may be shared, thus



Leveling Effect of a Solvent—When the strong acids such as $HClO_4$, H_2SO_4, HCl or HNO_3 are dissolved in water the solutions—if they are of identical normality and are not too concentrated—all have about the same hydrogen-ion concentration, indicating the acids to be of about the same strength. The reason for this is that each one of the acids undergoes practically complete protolysis in water.

$$\underset{\text{acid}_1}{HCl} + \underset{\text{base}_2}{H_2O} \rightarrow \underset{\text{base}_1}{Cl^-} + \underset{\text{acid}_2}{H_3O^+}$$

This phenomenon, called the leveling effect of water, occurs whenever the added acid is stronger than the hydronium ion. Such a reaction manifests the tendency of proton-transfer reactions to proceed spontaneously in the direction of forming a weaker acid or weaker base.

Since the strongest acid that can exist in an amphiprotic solvent is the conjugate acid form of the solvent, any stronger acid will undergo protolysis to the weaker solvent acid. Since $HClO_4$, H_2SO_4, HCl or HNO_3 are all stronger acids than the hydronium ion, they are converted in water to the hydronium ion.

When the strong bases sodium hydride, sodium amide or sodium ethoxide are dissolved in water, each reacts with water to form sodium hydroxide. These reactions illustrate the leveling effect of water on bases. Since the hydroxide ion is the strongest base which can exist in water, any base stronger than the hydroxide ion undergoes protolysis to hydroxide.

Intrinsic differences in the acidity of acids become evident if they are dissolved in a relatively poor proton acceptor such as anhydrous acetic acid. Perchloric acid ($HClO_4$), a strong acid, undergoes practically complete reaction with acetic acid

$$\underset{\substack{\text{acid}_1\\(\text{strong})}}{HClO_4} + \underset{\substack{\text{base}_2\\(\text{strong})}}{CH_3COOH} \rightarrow \underset{\substack{\text{base}_1\\(\text{weak})}}{ClO_4^-} + \underset{\substack{\text{acid}_2\\(\text{weak})}}{CH_3COOH_2^+}$$

but sulfuric acid and hydrochloric acid behave as weak acids. It is because perchloric acid is a very strong acid when dissolved in glacial acetic acid that it has found many important applications in analytical chemistry as a titrant for a variety of substances which behave as bases in acetic acid. Because of its ability to differentiate the acidity of various acids, it is called a *differentiating solvent for acids*; this property results from its relatively weak proton-acceptor tendency. A solvent that differentiates basicity of different hases must bave a weak proton-donor tendency; it is called a *differentiating solvent for bases*. Typical of solvents in this category is liquid ammonia. Solvents that have both weak proton-donor and proton-acceptor tendencies are called *aprotic solvents* and may serve as differentiating solvents for both acids and bases; they have little if any action on solutes and serve mainly as inert dispersion media for the solutes. Useful aprotic solvents are benzene, toluene or hexane.

Ionization of Acids and Bases—Acids and bases commonly are classified as strong or weak acids and strong or

weak bases depending on whether they are ionized extensively or slightly in aqueous solutions. If, for example, 1 *N* aqueous solutions of hydrochloric acid and acetic acid are compared, it is found that the former is a better conductor of electricity, reacts much more readily with metals, catalyzes certain reactions more efficiently and possesses a more acid taste than the latter. Both solutions, however, will neutralize identical amounts of alkali. A similar comparison of 1 *N* solutions of sodium hydroxide and ammonia reveals the former to be more "active" than the latter, although both solutions will neutralize identical quantities of acid.

The differences in the properties of the two acids is attributed to differences in the concentration of hydrogen (more accurately hydronium) ion, the hydrochloric acid being ionized to a greater extent and, therefore, containing a higher concentration of hydrogen ion than acetic acid. Similarly, most of the differences between the sodium hydroxide and ammonia solutions are attributed to the higher hydroxyl-ion concentration in the former.

The ionization of incompletely ionized acids may be considered a reversible reaction of the type

$$HA \rightleftharpoons H^+ + A^-$$

where HA is the molecular acid and A^- is its anion. An equilibrium expression based on the law of mass action may be applied to the reaction

$$K_a = \frac{[H^+][A^-]}{[HA]} \tag{14}$$

where K_a is the ionization or dissociation constant, and the brackets signify concentration. For any given acid in any specified solvent and at any constant temperature, K_a remains relatively constant as the concentration of acid is varied, provided the acid is weakly ionized. With increasingly stronger acids, however, progressively larger deviations occur.

Although the strength of an acid commonly is measured in terms of the ionization or dissociation constant defined in Eq 14, the process of ionization probably is never as simple as shown above. A proton simply will not detach itself from one molecule unless it is accepted simultaneously by another molecule. When an acid is dissolved in water, the latter acts as a base, accepting a proton (Brønsted's definition of a base) by donating a share in a pair of electrons (Lewis' definition of a base). This reaction may be written

$$\underset{\text{acid}_1}{HA} + \underset{\text{base}_2}{H_2O} \rightleftharpoons \underset{\text{base}_1}{A^-} + \underset{\text{acid}_2}{H_3O^+}$$

Application of the law of mass action to this reaction gives

$$K = \frac{[H_3O^+][A^-]}{[HA][H_2O]} \tag{15}$$

since $[H_2O]$ is a constant this equation may be written

$$K_a = \frac{[H_3O^+][A^-]}{[HA]} \tag{16}$$

This equation is identical with Eq 14 because $[H_3O^+]$ is numerically equal to $[H^+]$.

Acids which are capable of donating more than one proton are termed *polyprotic*. The ionization of a polyprotic acid occurs in stages and can be illustrated by considering the equilibria involved in the ionization of phosphoric acid

$$H_3PO_4 + H_2O \rightleftharpoons H_2PO_4^- + H_3O^+$$

$$H_2PO_4^- + H_2O \rightleftharpoons HPO_4^{2-} + H_3O^+$$

$$HPO_4^{2-} + H_2O \rightleftharpoons PO_4^{3-} + H_3O^+$$

Application of the law of mass action to this series of reactions gives

$$K_1 = \frac{[H_2PO_4^-][H_3O^+]}{[H_3PO_4]} \tag{17}$$

$$K_2 = \frac{[HPO_4^{2-}][H_3O^+]}{[H_2PO_4^-]} \tag{18}$$

$$K_3 = \frac{[PO_4^{3-}][H_3O^+]}{[HPO_4^{2-}]} \tag{19}$$

If the three expressions for the ionization constants are multiplied together, an overall ionization, K, can be obtained

$$K = K_1K_2K_3 = \frac{[PO_4^{3-}][H_3O^+]^3}{[H_3PO_4]} \tag{20}$$

Each of the successive ionizations is suppressed by the hydronium ion formed from preceding stages according to Le Chatelier's principle. The successive dissociation constants always decrease in value, since successive protons must be removed from species that are always more negatively charged. This can be seen from the data in Table IV, in which K_1 for phosphoric acid is approximately 100,000 times greater than K_2, which is in turn approximately 100,000 times greater than K_3. Although successive dissociation constants are always smaller, the difference is not always as great as it is for phosphoric acid. Tartaric acid, for example, has $K_1 = 9.12 \times 10^{-4}$ and $K_2 = 4.27 \times 10^{-5}$.

Ionization of a base can be illustrated by using the specific substance NH_3 for an example. According to Brønsted and Lewis, when the base NH_3 is dissolved in water, the latter acts as an acid, donating a proton to NH_3, which accepts it by offering a share in a pair of electrons on the nitrogen atom. This reaction is written

$$\underset{\text{base}}{NH_3} + \underset{\text{acid}}{H_2O} \rightleftharpoons NH_4^+ + OH^-$$

The equilibrium expression for this reaction is

$$K = \frac{[NH_4^+][OH^-]}{[NH_3][H_2O]} \tag{21}$$

with $[H_2O]$ constant this expression may be written

$$K_b = \frac{[NH_4^+][OH^-]}{[NH_3]} \tag{22}$$

Ionization of Water—Although it is a poor conductor of electricity, pure water does ionize through a process known as *autoprotolysis*, in the following manner

$$2H_2O \rightleftharpoons H_3O^+ + OH^-$$

Application of the law of mass action to this reaction gives

$$K = \frac{[H_3O^+][OH^-]}{[H_2O]^2} \tag{23}$$

where K is the equilibrium constant for the reaction. Since the concentration of H_2O (molecular water) is very much greater than either the hydronium-ion or hydroxyl-ion concentrations, it can be considered to be constant and can be combined with K to give a new constant, K_w, known as the *ion product* of water, and Eq 23 becomes

$$K_w = [H_3O^+][OH^-] \tag{24}$$

The numerical value of K_w varies with temperature; at 25° it is approximately equal to 1×10^{-14}.

Since the autoprotolysis of pure water yields one hydronium ion for each hydroxyl ion produced, $[H_3O^+]$ must be equal to $[OH^-]$. At 25° each has a value of 1×10^{-7} moles/

Table IV—Dissociation Constants in Water at 25°

Substance		K
Weak Acids		
Acetic		1.75×10^{-5}
Acetylsalicylic		3.27×10^{-4}
Barbital		1.23×10^{-8}
Barbituric		1.05×10^{-4}
Benzoic		6.30×10^{-5}
Benzyl penicillin		1.74×10^{-3}
Boric	K_1	5.8×10^{-10}
Caffeine		1×10^{-14}
Carbonic	K_1	4.31×10^{-7}
	K_2	4.7×10^{-11}
Citric ($1H_2O$)	K_1	7.0×10^{-4}
	K_2	1.8×10^{-5}
	K_3	4.0×10^{-7}
Dichloroacetic		5×10^{-2}
Ethylenediaminetetra-	K_1	1×10^{-2}
acetic acid (EDTA)	K_2	2.14×10^{-3}
	K_3	6.92×10^{-7}
	K_4	5.5×10^{-11}
Formic		1.77×10^{-4}
Glycerophosphoric	K_1	3.4×10^{-2}
	K_2	6.4×10^{-7}
Glycine	K_1	4.5×10^{-3}
	K_2	1.7×10^{-10}
Lactic		1.39×10^{-4}
Mandelic		4.29×10^{-4}
Monochloroacetic		1.4×10^{-3}
Oxalic ($2H_2O$)	K_1	5.5×10^{-2}
	K_2	5.3×10^{-5}
Phenobarbital		3.9×10^{-8}
Phenol		1×10^{-10}
Phosphoric	K_1	7.5×10^{-3}
	K_2	6.2×10^{-8}
	K_3	2.1×10^{-13}
Picric		4.2×10^{-1}
Propionic		1.34×10^{-5}
Saccharin		2.5×10^{-2}
Salicylic		1.06×10^{-3}
Succinic	K_1	6.4×10^{-5}
	K_2	2.3×10^{-6}
Sulfadiazine		3.3×10^{-7}
Sulfamerazine		8.7×10^{-8}
Sulfapyridine		3.6×10^{-9}
Sulfathiazole		7.6×10^{-8}
Tartaric	K_1	9.6×10^{-4}
	K_2	4.4×10^{-5}
Trichloroacetic		1.3×10^{-1}
Weak Bases		
Acetanilide		4.1×10^{-14} (40°)
Ammonia		1.74×10^{-5}
Apomorphine		1.0×10^{-7}
Atropine		4.5×10^{-5}
Benzocaine		6.0×10^{-12}
Caffeine		4.1×10^{-14} (40°)
Cocaine		2.6×10^{-6}
Codeine		9×10^{-7}
Ephedrine		2.3×10^{-5}
Morphine		7.4×10^{-7}
Papaverine		8×10^{-9}
Physostigmine	K_1	7.6×10^{-7}
	K_2	5.7×10^{-13}
Pilocarpine	K_1	7×10^{-8}
	K_2	2×10^{-13}
Procaine		7×10^{-6}
Pyridine		1.4×10^{-9}
Quinine	K_1	1.0×10^{-6}
	K_2	1.3×10^{-10}
Reserpine		4×10^{-8}
Strychnine	K_1	1×10^{-6}
	K_2	2×10^{-12}
Theobromine		4.8×10^{-14} (40°)
Thiourea		1.1×10^{-15}
Urea		1.5×10^{-14}

liter ($1 \times 10^{-7} \times 1 \times 10^{-7} = K_w = 1 \times 10^{-14}$). A solution in which $[H_3O^+]$ is equal to $[OH^-]$ is termed a *neutral* solution.

If an acid is added to water, the hydronium-ion concentration will be increased and the equilibrium between hydronium and hydroxyl ions will be disturbed *momentarily*. To restore equilibrium, some of the hydroxyl ions, originally present in the water, will combine with a *part* of the added hydronium ions to form nonionized water molecules, until the product of the concentrations of the two ions has been reduced to 10^{-14}. When equilibrium again is restored, the concentrations of the two ions no longer will be equal. If, for example, the hydronium-ion concentration is 1×10^{-3} N when equilibrium is established, the concentration of hydroxyl ion will be 1×10^{-11} (the product of the two concentrations being equal to 10^{-14}). Since $[H_3O^+]$ is much greater than $[OH^-]$, the solution is said to be *acid* or *acidic*.

In a similar manner, the addition of an alkali to pure water momentarily disturbs the equilibrium between hydronium and hydroxyl ions. To restore equilibrium, some of the hydronium ions originally present in the water will combine with part of the added hydroxyl ions to form nonionized water molecules. The process continues until the product of the hydronium and hydroxyl ion concentrations again is equal to 10^{-14}. Assuming that the final hydroxyl-ion concentration is 1×10^{-4} N, the concentration of hydronium ion in the solution will be 1×10^{-10}. Since $[OH^-]$ is much greater than $[H_3O^+]$, the solution is said to be *basic* or *alkaline*.

Relationship of K_a and K_b—A particularly interesting and useful relationship between the strength of an acid and its conjugate base, or a base and its conjugate acid, exists. For illustration, consider the strength of the base NH_3 and its conjugate acid NH_4^+, in water. The behavior of NH_3 as a base is expressed by

$$NH_3 + H_2O \rightleftharpoons NH_4^+ + OH^-$$

for which the equilibrium, as formulated earlier is

$$K_b = \frac{[NH_4^+][OH^-]}{[NH_3]} \tag{25}$$

The behavior of NH_4^+ as an acid is represented by

$$NH_4^+ + H_2O \rightleftharpoons NH_3 + H_3O^+$$

the equilibrium constant for which is

$$K_a = \frac{[NH_3][H_3O^+]}{[NH_4^+]} \tag{26}$$

Multiplying Eqs 25 and 26

$$K_aK_b = \frac{[NH_3][H_3O^+][NH_4^+][OH^-]}{[NH_4^+][NH_3]} \tag{27}$$

It is obvious that the product

$$K_w = K_a.K_b \tag{28}$$

where K_w is the ion product of water as defined in Eq 24.

The utility of this relationship, which is a general one for any conjugate acid–base pair, is evident from the following deductions: (1) the strength of an acid may be expressed in terms either of the K_a or the K_b of its conjugate base, or *vice versa;* (2) the K_a of an acid may be calculated if the K_b of its conjugate base is known, or *vice versa* and (3) the stronger an acid is, the weaker its conjugate base, or *vice versa*.

Bases which are capable of interacting with more than one proton are termed *polyacidic*, and can be illustrated by

$$PO_4^{3-} + H_2O \rightleftharpoons HPO_4^{2-} + OH^-$$

$$HPO_4^{2-} + H_2O \rightleftharpoons H_2PO_4^- + OH^-$$

$$H_2PO_4^- + H_2O \rightleftharpoons H_3PO_4 + OH^-$$

Applying the law of mass action to this series of reactions, and using the concepts outlined in Eqs 25 to 28, it becomes obvious that the relationship between the various K_a and K_b values for phosphoric acid are

$$K_w = K_{a1} \times K_{b3} = K_{a2} \times K_{b2} = K_{a3} \times K_{b1} \qquad (29)$$

where K_{a1}, K_{a2} and K_{a3} refer to the equilibria given by Eqs 17, 18 and 19, respectively; K_{b1}, K_{b2} and K_{b3} refer to the reaction of PO_4^{3-}, HPO_4^{2-} and $H_2PO_4^-$, respectively, with water.

Electronegativity and Dissociation Constants—Table IV gives the dissociation constants of several weak acids and weak bases, in water, at 25°. As pointed out previously, strong acids and strong bases do not obey the law of mass action, so that dissociation constants cannot be formulated for these strong electrolytes.

From an inspection of this table it is evident that great variations occur in the strength of weak acids and weak bases. The effect of various substituents on the strength of acids and bases depends on the electronegativity of the substituent atom or radical. For example, the substitution of one chlorine atom into the molecule of acetic acid increases the degree of ionization of the acid. Substitution of two chlorine atoms further increases the degree of ionization, and introduction of three chlorine atoms produces a still stronger acid. Acetic acid ionizes primarily because the oxygen atom adjacent to the hydrogen atom of the carboxyl group has a stronger affinity for electrons than has the hydrogen atom. Thus, when acetic acid is dissolved in water the polar molecules of the latter have a stronger affinity for the hydrogen of acetic acid than the latter. The acetic acid ionizes as a consequence of this difference in affinities. When an atom of chlorine is introduced into the acetic acid molecule, forming $ClCH_2COOH$, the electrons in the molecule are attracted very strongly to the chlorine because of its relatively high electronegativity, the bond between the hydrogen and the oxygen in the carboxyl group thereby weakened and the degree of ionization increased. Introduction of two, or three chlorine atoms weakens the bond further and increases the strength of the acid. On the other hand, substitution of chlorine into the molecule of ammonia reduces the strength of the base because of its decreased affinity for the hydrogen ion.

Ionic Strength and Dissociation Constants—Most solutions of pharmaceutical interest are in a concentration range such that the ionic strength of the solution may have a marked effect on ionic equilibria and observed dissociation constants. One method of correcting dissociation constants for solutions with an ionic strength up to about 0.3 is to calculate an apparent dissociation constant, pK_a' as

$$pK_a' = pK_a + \frac{0.51\,(2Z-1)\,\sqrt{\mu}}{1+\sqrt{\mu}} \qquad (30)$$

in which pK_a is the tabulated thermodynamic dissociation constant, Z is the charge on the acid and μ is the ionic strength.

Example—Calculate pK_2' for succinic acid at an ionic strength of 0.1. Assume that pK_2 is 5.63. The charge on the acid species is −1.

$$pK_2' = 5.63 + \frac{0.51\,(-2-1)\,\sqrt{0.1}}{1+\sqrt{0.1}}$$

$$= 5.63 - 0.37 = 5.26$$

Determination of Dissociation Constants—Although the dissociation constant of a weak acid or base can be obtained in a wide variety of ways including conductivity measurements, or ultraviolet or visible absorption spectrometry, the most widely used method is potentiometric pH measurement (see *Potentiometry*, page 244). The simplest method involving potentiometric pH measurement is based on the measurement of the hydronium-ion concentration of a solution containing equimolar concentrations of the acid and a strong-base salt of the acid. The principle of this method is evident from an inspection of Eq 16; when equimolar concentrations of HA (the acid) and A^- (the salt) are present, the dissociation constant, K_a, numerically is equal to the hydronium-ion concentration (also, the pK_a of the acid is equal to the pH of the solution). Although this method is simple and rapid, the dissociation constant obtained is not sufficiently accurate for many purposes.

In order to obtain the dissociation constant of a weak acid with a high degree of accuracy and precision, a dilute solution of the acid (about 10^{-3} to 10^{-4} *M*) is titrated with a strong base, and the pH of the solution taken after each addition of base. The resulting data can be handled in a wide variety of ways, perhaps the best of which is the method proposed by Benet and Goyan.[1] The proton balance equation for a weak acid, HA, being titrated with a strong base such as KOH, would be

$$[K^+] + [H_3O^+] = [OH^-] + [A^-] \qquad (31)$$

in which $[K^+]$ is the concentration of the base added. Equation 31 can be rearranged to give

$$Z = [A^-] = [K^+] + [H_3O^+] - [OH^-] \qquad (32)$$

When a weak monoprotic acid is added to water, it can exist in the unionized form, HA, and in the ionized form, A^-. After equilibrium is established, the sum of the concentrations of both species must be equal to C_a, the stoichiometric (added) concentration of acid or

$$C_a = [HA] + [A^-] = [HA] + Z \qquad (33)$$

The term, [HA], can be replaced using Eq 16 to give

$$C_a = \frac{[H_3O^+]Z}{K_a} + Z \qquad (34)$$

which can be rearranged to

$$Z = C_a - \frac{Z[H_3O^+]}{K_a} \qquad (35)$$

According to Eq 35, if Z, which is obtained from the experimental data using Eq 32, is plotted versus the terms $Z[H_3O^+]$, a straight line results with a slope equal to $1/K_a$, and an intercept equal to C_a. In addition to obtaining an accurate estimate for the dissociation constant, the stoichiometric concentration of the substance being titrated is obtained also. This is of importance when the substance being titrated cannot he purified, or has an unknown degree of solvation. Similar equations can be developed for ohtaining the dissociation constant for a weak base.[1]

The dissociation constants for diprotic acids can be obtained by defining P as the average number of protons dissociated per mole of acid or

$$P = Z/C_a \qquad (36)$$

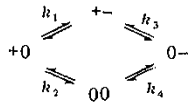
and

$$\frac{[H_3O^+]^2P}{(2-P)} = K_1K_2 + \frac{K_1[H_3O^+]\,(1-P)}{(2-P)} \qquad (37)$$

A plot of Eq 37 should yield a straight line with a slope equal to K_1 and an intercept of K_2; dividing the intercept by the slope yields K_2.

Micro Dissociation Constants—The dissociation constants for polyprotic acids, as determined hy potentiometric titration, are generally known as macro, or titration con-

stants. Since it is known that carboxyl groups are stronger acids than protonated amino groups, there is no difficulty in assigning K_1 and K_2, as determined by Eq 37 to the carboxyl and amino groups, respectively, of a substance such as glycine hydrochloride. In other chemicals or drugs such as phenylpropanolamine, in which the two acidic groups are the phenolic and the protonated amino group, the assignment of dissociation constants is more difficult. This is because, in general, both groups have dissociation constants of equal magnitude. Thus, there will be two ways of losing the first proton and two ways of losing the second, resulting in four possible species in solution. This can be illustrated using the convention of assigning a + to a positively charged group, a 0 to an uncharged group and a − to a negatively charged group. Thus, +0 would represent the fully protonated phenylpropanolamine, +− the dipolar ion, 00 the uncharged molecule and 0−, the anion. The total ionization scheme, therefore, can be written



The micro constants are related to the macro constants as

$$K_1 = k_1 + k_2 \tag{38}$$

$$K_1K_2 = k_1k_3 = k_2k_4 \tag{39}$$

It can be seen from Eq 38 that unless k_1 or k_2 is very much smaller than the other, the observed macro constant is a composite of the two and cannot he assigned to one or the other acidic group in a nonambiguous way.

Methods for determining k_1 are given by Riegelman[2] and Niebergall *et al.*[3] Once k_1, K_1 and K_2 have been determined, all of the other micro constants can be obtained from Eqs 38 and 39.

pH

The numerical values of hydronium-ion concentration may vary enormously; for a normal solution of a strong acid the value is nearly 1, while for a normal solution of a strong base it is approximately 1×10^{-14}; ie, a variation of 100,000,000,000,000 between these two limits. Because of the inconvenience of dealing with such large numbers, Sørenson, in 1909, proposed that hydronium-ion concentration be expressed in terms of the logarithm (log) of its reciprocal. To this value he assigned the symbol pH. Mathematically it is written

$$\text{pH} = \log \frac{1}{[H_3O^+]} \tag{40}$$

and since the logarithm of 1 is zero, the equation also may be written

$$\text{pH} = -\log[H_3O^+] \tag{41}$$

from which it is evident that pH also may be defined as the negative logarithm of tbe hydronium-ion concentration. In general, this type of notation is used to indicate the negative logarithm of the term that is preceded by the "p," which gives rise to the following

$$\text{pOH} = -\log [OH^-] \tag{42}$$

$$pK = -\log K \tag{43}$$

Thus, taking logarithms of Eqs 26 and 28 gives

$$pK_a + pK_b = pK_w \tag{44}$$

$$\text{pH} + \text{pOH} = pK_w \tag{45}$$

The relationship of pH to hydronium-ion and hydroxyl-ion concentrations may be seen in Table V.

The following examples illustrate the conversion from exponential to "p" notation

1. Calculate the pH corresponding to a hydronium-ion concentration of 1×10^{-4} g-ion/L.

Solution:

$$\text{pH} = \log \frac{1}{1 \times 10^{-4}}$$

$= \log 10{,}000$ or $\log (1 \times 10^{+4})$

$\log (1 \times 10^{+4}) = +4$

$\text{pH} = 4$

Table V—Hydronium-Ion and Hydroxyl-Ion Concentrations

	pH	Normality in Terms of Hydronium Ion	Normality in Terms of Hydroxyl Ion
	0	1	10^{-14}
	1	10^{-1}	10^{-13}
Increasing acidity	2	10^{-2}	10^{-12}
	3	10^{-3}	10^{-11}
	4	10^{-4}	10^{-10}
	5	10^{-5}	10^{-9}
	6	10^{-6}	10^{-8}
Neutral point	7	10^{-7}	10^{-7}
	8	10^{-8}	10^{-6}
	9	10^{-9}	10^{-5}
Increasing alkalinity	10	10^{-10}	10^{-4}
	11	10^{-11}	10^{-3}
	12	10^{-12}	10^{-2}
	13	10^{-13}	10^{-1}
	14	10^{-14}	1

2. Calculate the pH corresponding to a hydronium ion-concentration of 0.000036 *N* (or g-ion/L) (*Note*—This more frequently is written as a number multiplied by a power of 10, thus, 3.6×10^{-5} for 0.000036.)

Solution

$$\text{pH} = \log \frac{1}{3.6 \times 10^{-5}}$$

$= \log 28{,}000$ or $\log (2.8 \times 10^{+4})$

$\log (2.8 \times 10^{+4}) = \log 2.8 + \log 10^{+4}$

$\log 2.8 = +0.44$

$\log 10^{+4} = \underline{+4.00}$

$\text{pH} = 4.44$

This problem may also be solved as follows

$\text{pH} = -\log (3.6 \times 10^{-5})$

$\log 3.6 = +0.56$

$\log 10^{-5} = \underline{-5.00}$

$= -4.44 = \log (3.6 \times 10^{-5})$

$\text{pH} = -(-4.44) = +4.44 = 4.44$

The following examples illustrate the conversion of "p" notation to exponential notation:

1. Calculate the hydronium-ion concentration corresponding to a pH of 4.44.

Solution

$$\text{pH} = \log \frac{1}{[H_3O^+]}$$

$$4.44 = \log \frac{1}{[H_3O^+]}$$

$$\frac{1}{[H_3O^+]} = \text{antilog of } 4.44 = 28{,}000 \text{ (rounded off)}$$

$$[H_3O^+] = \frac{1}{28{,}000} = 0.000036 \text{ or } 3.6 \times 10^{-5}$$

This calculation also may be made as

$$+4.44 = -\log [H_3O^+]$$

$$\text{or } -4.44 = +\log [H_3O^+]$$

In finding the antilog of −4.44 it should be kept in mind that the *mantissa* (the number to the right of the decimal point) of a log to the base 10 (the common or Briggsian logarithm base) is *always positive* but that the *characteristic* (the number to the left of the decimal point) may be *positive* or *negative*. As the entire log −4.44 is negative, it is obvious that one cannot look up the antilog of −0.44. However, the number −4.44 also may be written (−5.00 + 0.56) or, as more often written, $\bar{5}.56$, where the bar across the characteristic indicates that it alone is negative, while the rest of the number is positive. Looking up the antilog of 0.56 it is found to be 3.6 and, as the antilog of −5.00 is 10^{-5}, it follows that the hydronium-ion concentration must be 3.6×10^{-5} moles/L.

2. Calculate the hydronium-ion concentration corresponding to a pH of 10.17.

Solution

$$10.17 = -\log [H_3O^+]$$

$$-10.17 = \log [H_3O^+]$$

$$-10.17 = (-11.00 + 0.83) = \bar{1}\bar{1}.83$$

The antilog of 0.83 = 6.8
The antilog of −11.00 = 10^{-11}
The hydronium-ion concentration is therefore 6.8×10^{-11} moles/L.

In the section on *Ionization of Water* it was shown that the hydronium-ion concentration of pure water, at 25°, is 1×10^{-7} *N*, corresponding to a pH of 7.* This figure, therefore, is designated as the neutral point and all values below a pH of 7 represent acidity; the smaller the number, the greater the acidity. Values above 7 represent alkalinity; the larger the number, the greater the alkalinity. The pH scale usually runs from 0 to 14, but mathematically there is no reason why negative numbers or numbers above 14 should not be used. In practice, however, such values are never encountered because solutions which might be expected to have such values are too concentrated to be ionized extensively or the interionic attraction is so great as to materially reduce ionic activity.

It should be emphasized strongly that the generalizations stated concerning neutrality, acidity and alkalinity hold exactly only when (1) the solvent is water, (2) the temperature is 25° and (3) there are no other factors to cause deviation from the simply formulated equilibria underlying the definition of pH given in the preceding discussion.

Species Concentration

When a weak acid, H_nA, is added to water, $n + 1$ species, including the unionized acid, can exist. After equilibrium is established, the sum of the concentrations of all species must be equal to C_a, the stoichiometric (added) concentration of acid. Thus, for a triprotic acid H_3A

$$C_a = [H_3A] + [H_2A^-] + [HA^{2-}] + [A^{3-}] \tag{46}$$

In addition, the concentrations of all acidic and basic species in solution vary with pH, and can be represented solely in terms of equilibrium constants and the hydronium-ion concentration. These relationships may be expressed as

$$[H_nA] = [H_3O^+]^n C_a/D \tag{47}$$

$$[H_{n-j}A^{-j}] = [H_3O^+]^{n-j} K_1 \ldots K_j C_a/D \tag{48}$$

in which n represents the total number of dissociable hydrogens in the parent acid, j is the number of protons dissociated, C_a is the stoichiometric concentration of acid and K represents the acid dissociation constants. The term D is a power series in $[H_3O^+]$ and K, starting with $[H_3O^+]$ raised to the nth power. The last term is the product of all the dissociation constants. The intermediate terms can be generated from the last term by substituting $[H_3O^+]$ for K_n to obtain the next-to-last term, then substituting $[H_3O^+]$ for K_{n-1} to obtain the next term, etc, until the first term is reached. The following examples show the denominator, D, to be used for various types of acids

$$H_3A\text{: } D = [H_3O^+]^3 + K_1[H_3O^+]^2 + K_1K_2[H_3O^+] + K_1K_2K_3 \tag{49}$$

$$H_2A^-\text{: } D = [H_3O^+]^2 + K_1[H_3O^+] + K_1K_2 \tag{50}$$

$$HA^{2-}\text{: } D = [H_3O^+] + K_a \tag{51}$$

The numerator, in all instances, is C_a multiplied by the term from the denominator that has $[H_3O^+]$ raised to the $n - j$ power. Thus, for diprotic acids such as carbonic, succinic, tartaric, etc

$$[H_2A] = \frac{[H_3O^+]^2 C_a}{[H_3O^+]^2 + K_1[H_3O^+] + K_1K_2} \tag{52}$$

$$[HA^-] = \frac{K_1[H_3O^+]C_a}{[H_3O^+]^2 + K_1[H_3O^+] + K_1K_2} \tag{53}$$

$$[A^{2-}] = \frac{K_1K_2C_a}{[H_3O^+]^2 + K_1[H_3O^+] + K_1K_2} \tag{54}$$

Example—Calculate the concentrations of all succinic acid species in a 1.0×10^{-3} *M* solution of succinic acid at pH 6.0. Assume that $K_1 = 6.4 \times 10^{-5}$ and $K_2 = 2.3 \times 10^{-6}$.

Eqs 52–54 have the same denominator, D, which can be calculated as

$$D = [H_3O^+]^2 + K_1[H_3O^+] + K_1K_2$$

$$= 1.0 \times 10^{-12} + 6.4 \times 10^{-5} \times 1.0 \times 10^{-6} + 6.4 \times 10^{-5} \times 2.3 \times 10^{-6}$$

$$= 1.0 \times 10^{-12} + 6.4 \times 10^{-11} + 14.7 \times 10^{-11}$$

$$= 21.2 \times 10^{-11}$$

Therefore

$$[H_2A] = \frac{[H_3O^+]^2 C_a}{D} = \frac{1.0 \times 10^{-12} \times 1.0 \times 10^{-3}}{21.2 \times 10^{-11}} = 4.7 \times 10^{-6}\, M$$

* The pH of the purest water obtainable, so-called "conductivity" water, is 7.0 when the measurement is carefully made under conditions to exclude carbon dioxide and prevent errors inherent in the measuring technique (such as acidity or alkalinity of the indicator). Upon agitating this water in the presence of carbon dioxide in the atmosphere (equilibrium water) the value drops rapidly to 5.7, which is the pH of nearly all distilled waters that have been exposed to the atmosphere for even a short time.

$$[HA^-] = \frac{K_1[H_3O^+]C_a}{D} =$$

$$\frac{6.4 \times 10^{-11} \times 1.0 \times 10^{-3}}{21.2 \times 10^{-11}} = 3.0 \times 10^{-4}\ M$$

$$[A^{2-}] = \frac{K_1K_2C_a}{D} =$$

$$\frac{14.7 \times 10^{-11} \times 1.0 \times 10^{-3}}{21.2 \times 10^{-11}} = 6.9 \times 10^{-4}\ M$$

Proton Balance Equation

In the Brønsted–Lowry system the total number of protons released by acidic species must equal the total number of protons consumed by basic species. This results in a very useful relationship known as the proton balance equation (PBE), in which the sum of the concentration terms for species that form by proton consumption is equated to the sum of the concentration terms for species that are formed by the release of protons. The PBE forms the basis of a unified approach to pH calculations, since it is an exact accounting of all proton transfers occurring in solution.

When HCl is added to water, for example, it dissociates yielding one Cl^- for each proton released. Thus, Cl^- is a species formed by the release of a proton. In the same solution, and actually in all aqueous solutions

$$2H_2O \rightleftharpoons H_3O^+ + OH^-$$

where H_3O^+ is formed by proton consumption and OH^- is formed by proton release. Thus, the PBE is

$$[H_3O^+] = [OH^-] + [Cl^-] \tag{55}$$

In general, the PBE can be formed in the following manner

1. Start with the species added to water.
2. Place all species that can form when protons are released on the right side of the equation.
3. Place all species that can form when protons are consumed on the left side of the equation.
4. Add $[H_3O^+]$ to the left side of the equation and $[OH^-]$ to the right side of the equation. These result from the interaction of two molecules of water as shown above.

Example—When H_3PO_4 is added to water, the species $H_2PO_4^-$ forms with the release of one proton, HPO_4^{2-} forms with the release of two protons and PO_4^{3-} forms with the release of three protons to give the following PBE

$$[H_3O^+] = [OH^-] + [H_2PO_4^-] + 2[HPO_4^{2-}] + 3[PO_4^{3-}] \tag{56}$$

Example—When Na_2HPO_4 is added to water, it dissociates into two Na^+ and one HPO_4^{2-}. The sodium ion is neglected in the PBE since it is not formed from the release or consumption of protons. The species HPO_4^{2-}, however, may react with water to give $H_2PO_4^-$ with the consumption of one proton, H_3PO_4 with the consumption of two protons, and PO_4^{3-} with the release of one proton to give the following PBE

$$[H_3O^+] + [H_2PO_4^-] + 2[H_3PO_4] = [OH^-] + [PO_4^{3-}] \tag{57}$$

Calculations

The pH of solutions of acids, bases and salts may be calculated using the concepts presented in the preceding sections.

Strong Acids or Bases

When a strong acid such as HCl is added to water, the following reactions occur

$$HCl + H_2O \rightarrow H_3O^+ + Cl^-$$

$$2H_2O \rightleftharpoons H_3O^+ + OH^-$$

The proton balance equation for this system would be

$$[H_3O^+] = [OH^-] + [Cl^-] \tag{58}$$

In most instances ($C_a > 4.5 \times 10^{-7}\ M$) the $[OH^-]$ would be negligible compared to the $[Cl^-]$ and the equation simplifies to

$$[H_3O^+] = [Cl^-] = C_a \tag{59}$$

Thus, the hydronium-ion concentration of a solution of a strong acid would be equal to the stoichiometric concentration of the acid. This would be anticipated, since strong acids generally are assumed to be 100% ionized.

The pH of a 0.005 *M* solution of HCl therefore is calculated as

$$pH = -\log 0.005 = 2.30$$

In a similar manner the hydroxyl-ion concentration for a solution of a strong base such as NaOH would be

$$[OH^-] = [Na^+] = C_b \tag{60}$$

and the pH of a 0.005 *M* solution of NaOH would be

$$pOH = -\log 0.005 = 2.30$$

$$pH = pK_w - pOH = 14.00 - 2.30 = 11.70$$

Weak Acids or Bases

If a weak acid, HA, is added to water, it will equilibrate with its conjugate base, A^-, as

$$HA + H_2O \rightleftharpoons H_3O^+ + A^-$$

Accounting for the ionization of water gives the following proton balance equation for this system

$$[H_3O^+] = [OH^-] + [A^-] \tag{61}$$

The concentration of A^- as a function of hydronium-ion concentration can be obtained as shown previously to give

$$[H_3O^+] = [OH^-] + \frac{K_aC_a}{[H_3O^+] + K_a} \tag{62}$$

Algebraic simplification yields

$$[H_3O^+] = K_a\frac{(C_a - [H_3O^+] + [OH^-])}{([H_3O^+] - [OH^-])} \tag{63}$$

In most instances for solutions of weak acids, $[H_3O^+] \gg [OH^-]$ and the equation simplifies to give

$$[H_3O^+]^2 + K_a[H_3O^+] - K_aC_a = 0 \tag{64}$$

This is a quadratic equation* which yields

$$[H_3O^+] = \frac{-K_a + \sqrt{K_a^2 + 4K_aC_a}}{2} \tag{65}$$

since $[H_3O^+]$ can never be negative. Furthermore, if $[H_3O^+]$ is less than 5% of C_a, Eq 64 is simplified further to give

* The general solution to a quadratic equation of the form

$$aX^2 + bX + c = 0$$

is

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$[H_3O^+] = \sqrt{K_a C_a} \tag{66}$$

It generally is preferable to use the simplest equation to calculate $[H_3O^+]$. However, when $[H_3O^+]$ is calculated, it must be compared to C_a in order to determine whether the assumption $C_a \gg [H_3O^+]$ is valid. If the assumption is not valid, the quadratic equation should be used.

Example—Calculate the pH of a 5.00×10^{-5} *M* solution of a weak acid having a $K_a = 1.90 \times 10^{-5}$.

$$[H_3O^+] = \sqrt{K_a C_a}$$
$$= 1.90 \times 10^{-5} \times 5.00 \times 10^{-5}$$
$$= 3.08 \times 10^{-5}\ M$$

Since C_a $[(5.00 \times 10^{-5}\ M)]$ is not much greater than $[H_3O^+]$, the quadratic equation (Eq 65) should be used

$$[H_3O^+] = \frac{-1.90 \times 10^{-5} + \sqrt{(1.90 \times 10^{-5})^2 + 4(1.90 \times 10^{-5} \times 5.00 \times 10^{-5})}}{2}$$
$$= 2.26 \times 10^{-5}\ M$$
$$\text{pH} = -\log (2.26 \times 10^{-5}) = 4.65$$

Note that the assumption $[H_3O^+] \gg [OH^-]$ is valid. The hydronium-ion concentration calculated from Eq 66 has a relative error of 36% when compared to the correct value obtained from Eq 65.

When a salt obtained from a strong acid and a weak base—e.g., ammonium chloride, morphine sulfate, pilocarpine hydrochloride, etc—is dissolved in water, it dissociates as

$$BH^+X^- \xrightarrow{H_2O} BH^+ + X^-$$

in which BH^+ is the protonated form of the base B, and X^- is the anion of a strong acid. Since X^- is the anion of a strong acid, it is too weak a base to undergo any further reaction with water. The protonated base, however, can act as a weak acid to give

$$BH^+ + H_2O \rightleftharpoons B + H_3O^+$$

Thus, Eqs 65 and 66 are valid, with C_a being equal to the concentration of the salt in solution. If K_a for the protonated base is not available, it can be obtained by dividing K_b for the base B, into K_w.

Example—Calculate the pH of a 0.026 *M* solution of ammonium chloride. Assume that K_b for ammonia is 1.74×10^{-5} and K_w is 1.00×10^{-14}.

$$K_a = \frac{K_w}{K_b} = \frac{1.00 \times 10^{-14}}{1.74 \times 10^{-5}} = 5.75 \times 10^{-10}$$
$$[H_3O^+] = \sqrt{K_a C_a}$$
$$= \sqrt{5.75 \times 10^{-10} \times 2.6 \times 10^{-2}}$$
$$= 3.87 \times 10^{-6}\ M$$
$$\text{pH} = -\log (3.87 \times 10^{-6}) = 5.41$$

Since C_a is much greater than $[H_3O^+]$ and $[H_3O^+]$ is much greater than $[OH^-]$, the assumptions are valid and the value calculated for pH is sufficiently accurate.

Weak Bases

When a weak base, B, is dissolved in water it ionizes to give the conjugate acid as

$$B + H_2O \rightleftharpoons BH^+ + OH^-$$

The proton balance equation for this system is

$$[BH^+] + [H_3O^+] = [OH^-] \tag{67}$$

Substituting $[BH^+]$ as a function of hydronium-ion concentration and simplifying, in the same manner as shown for a weak acid, gives

$$[OH^-] = K_b \frac{(C_b - [OH^-] + [H_3O^+])}{([OH^-] - [H_3O^+])} \tag{68}$$

If $[OH^-] \gg [H_3O^+]$, as is true generally

$$[OH^-]^2 + K_b[OH^-] - K_b C_b = 0 \tag{69}$$

which is a quadratic with the following solution

$$[OH^-] = \frac{-K_b + \sqrt{K_b^2 + 4K_b C_b}}{2} \tag{70}$$

If $C_b \gg [OH^-]$, the quadratic equation simplifies to

$$[OH^-] = \sqrt{K_b C_b} \tag{71}$$

Once $[OH^-]$ is calculated, it can be converted to pOH, which can be subtracted from pK_w to give pH.

Example—Calculate the pH of a 4.50×10^{-2} *M* solution of a weak base having $K_b = 2.00 \times 10^{-4}$. Assume that $K_w = 1.00 \times 10^{-14}$.

$$[OH^-] = \sqrt{K_b C_b}$$
$$= \sqrt{2.00 \times 10^{-4} \times 4.50 \times 10^{-2}}$$
$$= \sqrt{9.00 \times 10^{-6}} = 3.00 \times 10^{-3}\ M$$

Both assumptions are valid.

$$\text{pOH} = -\log 3.00 \times 10^{-3} = 2.52$$
$$\text{pH} = 14.00 - 2.52 = 11.48$$

When salts obtained from strong bases and weak acids (eg, sodium acetate, sodium sulfathiazole, sodium benzoate, etc) are dissolved in water, they dissociate as

$$Na^+A^- \xrightarrow{H_2O} Na^+ + A^-$$

in which A^- is the conjugate base of the weak acid, HA. The Na^+ undergoes no further reaction with water. The A^-, however, acts as a weak base to give

$$A^- + H_2O \rightleftharpoons HA + OH^-$$

Thus, Eqs 70 and 71 are valid, with C_b being equal to the concentration of the salt in solution. The value for K_b can be obtained by dividing K_a for the conjugate acid, HA, into K_w.

Example—Calculate the pH of a 0.05 *M* solution of sodium acetate. Assume that K_a for acetic acid $= 1.75 \times 10^{-5}$ and $K_w = 1.00 \times 10^{-14}$.

$$K_b = \frac{K_w}{K_a} = \frac{1.00 \times 10^{-14}}{1.75 \times 10^{-5}}$$
$$= 5.71 \times 10^{-10}$$
$$OH^- = \sqrt{K_b C_b} = \sqrt{5.71 \times 10^{-10} \times 5.0 \times 10^{-2}}$$
$$= 5.34 \times 10^{-6}\ M$$

Both assumptions are valid

$$\text{pOH} = -\log (5.34 \times 10^{-6}) = 5.27$$
$$\text{pH} = 14.00 - 5.27 = 8.73$$

Ampholytes

Substances such as $NaHCO_3$ and NaH_2PO_4 are termed *ampholytes*, and are capable of functioning both as acids and bases. When an ampholyte of the type NaHA is dissolved in water, the following series of reactions can occur

$$Na^+HA^- \xrightarrow{H_2O} Na^+ + HA^-$$

$$HA^- + H_2O \rightleftharpoons A^{2-} + H_3O^+$$

$$HA^- + H_2O \rightleftharpoons H_2A + OH^-$$

$$2H_2O \rightleftharpoons H_3O^+ + OH^-$$

The total proton balance equation (PBE) for the system is

$$[H_3O^+] + [H_2A] = [OH^-] + [A^{2-}] \tag{72}$$

Substituting both $[H_2A]$ and $[A^{2-}]$ as a function of $[H_3O^+]$ (see Eqs 52 and 54), yields

$$[H_3O^+] + \frac{[H_3O^+]^2 C_s}{[H_3O^+]^2 + K_1[H_3O^+] + K_1K_2} = \frac{K_w}{[H_3O^+]} + \frac{K_1K_2C_s}{[H_3O^+]^2 + K_1[H_3O^+] + K_1K_2}$$

This gives a fourth-order equation in $[H_3O^+]$, which can be simplified using certain judicious assumptions to

$$[H_3O^+] = \sqrt{\frac{K_1K_2C_s}{K_1 + C_s}} \tag{73}$$

In most instances, $C_s \gg K_1$ and the equation further simplifies to

$$[H_3O^+] = \sqrt{K_1K_2} \tag{74}$$

and $[H_3O^+]$ becomes independent of the concentration of the salt. A special property of ampholytes is that the concentration of the species HA^- is maximum at the pH corresponding to Eq 74.

When the simplest amino acid salt, glycine hydrochloride, is dissolved in water, it acts as a diprotic acid and ionizes as

$$^+NH_3CH_2COOH + H_2O \rightleftharpoons {}^+NH_3CH_2COO^- + H_3O^+$$

$$^+NH_3CH_2COO^- + H_2O \rightleftharpoons NH_2CH_2COO^- + H_3O^+$$

The form, $^+NH_3CH_2COO^-$, is an ampholyte since it also can act as a weak base

$$^+NH_3CH_2COO^- + H_2O \rightleftharpoons {}^+NH_3CH_2COOH + OH^-$$

This type of substance, which carries both a charged acidic and a charged basic moiety on the same molecule is termed a *zwitterion* and, since the two charges balance each other, the molecule acts essentially as a neutral molecule. The pH at which the *zwitterion* concentration is maximum is known as the isoelectric point, which can be calculated from Eq 74.

On the acid side of the isoelectric point, amino acids and proteins are cationic and incompatible with anionic materials such as the naturally occurring gums used as suspending and/or emulsifying agents. On the alkaline side of the isoelectric point, amino acids and proteins are anionic and incompatible with cationic materials such as benzalkonium chloride.

Salts of Weak Acids and Weak Bases

When a salt such as ammonium acetate (which is derived from a weak acid and a weak base) is dissolved in water, it undergoes the following reactions

$$BH^+A^- \xrightarrow{H_2O} BH^+ + A^-$$

$$BH^+ + H_2O \rightleftharpoons B + H_3O^+$$

$$A^- + H_2O \rightleftharpoons HA + OH^-$$

The total PBE for this system is

$$[H_3O^+] + [HA] = [OH^-] + [B] \tag{75}$$

Replacing [HA] and [B] as a function of $[H_3O^+]$, gives

$$[H_3O^+] + \frac{[H_3O^+]C_s}{[H_3O^+] + K_a} = [OH^-] + \frac{K_a'C_s}{[H_3O^+] + K_a'} \tag{76}$$

in which C_s is the concentration of salt, K_a is the ionization constant of the conjugate acid formed from the reaction between A^- and water and K_a' is the ionization constant for the protonated base, BH^+. In general, $[H_3O^+]$, $[OH^-]$, K_a and K_a' usually are smaller than C_s and the equation simplifies to

$$[H_3O^+] = \sqrt{K_aK_a'} \tag{77}$$

Example—Calculate the pH of a 0.01 *M* solution of ammonium acetate. The ammonium ion has a K_a equal to 5.75×10^{-10}, which represents K_a' in Eq 77. Acetic acid has a K_a of 1.75×10^{-5}, which represents K_a in Eq 77

$$[H_3O^+] = \sqrt{1.75 \times 10^{-5} \times 5.75 \times 10^{-10}}$$

$$= 1.05 \times 10^{-7}$$

$$pH = -\log (1.05 \times 10^{-7}) = 6.98$$

All of the assumptions are valid.

Buffers

The terms *buffer*, *buffer solution* and *buffered solution*, when used with reference to hydrogen-ion concentration or pH, refer to the ability of a system, particularly an aqueous solution, to resist a change of pH on adding acid or alkali, or on dilution with a solvent.

If an acid or base is added to water, the pH of the latter is changed markedly, for water has no ability to resist change of pH; it is completely devoid of buffer action. Even a very weak acid such as carbon dioxide changes the pH of water, decreasing it from 7 to 5.7 when the small concentration of carbon dioxide present in air is equilibrated with pure water. This extreme susceptibility of distilled water to a change of pH on adding very small amounts of acid or base is often of great concern in pharmaceutical operations. Solutions of neutral salts, such as sodium chloride, similarly lack ability to resist change of pH on adding acid or base; such solutions are called unbuffered.

Characteristic of buffered solutions, which undergo small changes of pH on addition of acid or base, is the presence either of a weak acid and a salt of the weak acid, or a weak base and a salt of the weak base. An example of the former system is acetic acid and sodium acetate; of the latter, ammonium hydroxide and ammonium chloride. From the proton concept of acids and bases discussed earlier, it is apparent that such buffer action involves a conjugate acid–base pair in the solution. It will be recalled that acetate ion is the conjugate base of acetic acid, and that ammonium ion is the conjugate acid of ammonia (the principal constituent of what commonly is called ammonium hydroxide).

The mechanism of action of the acetic acid–sodium acetate buffer pair is that the acid, which exists largely in molecular (nonionized) form, combines with hydroxyl ion that may be added to form acetate ion and water, thus

$$CH_3COOH + OH^- \rightarrow CH_3COO^- + H_2O$$

while the acetate ion, which is a base, combines with hydrogen (more exactly hydronium) ion that may be added to

form essentially nonionized acetic acid and water, represented as

$$CH_3COO^- + H_3O^+ \rightarrow CH_3COOH + H_2O$$

As will be illustrated later by an example, the change of pH is slight as long as the amount of hydronium or hydroxyl ion added does not exceed the capacity of the buffer system to neutralize it.

The ammonia–ammonium chloride pair functions as a buffer because the ammonia combines with hydronium ion that may be added to form ammonium ion and water, thus

$$NH_3 + H_3O^+ \rightarrow NH_4^+ + H_2O$$

Ammonium ion, which is an acid, combines with added hydroxyl ion to form ammonia and water, as

$$NH_4^+ + OH^- \rightarrow NH_3 + H_2O$$

Again, the change of pH is slight if the amount of added hydronium or hydroxyl ion is not in excess of the capacity of the system to neutralize it.

Besides these two general types of buffers, a third appears to exist. This is the buffer system composed of two salts, as monobasic potassium phosphate, KH_2PO_4, and dibasic potassium phosphate, K_2HPO_4. This is not, however, a new type of buffer; it is actually a weak-acid–conjugate-base buffer in which an ion, $H_2PO_4^-$, serves as the weak acid, and HPO_4^{2-} is its conjugate base. When hydroxyl ion is added to this buffer the following reaction takes place

$$H_2PO_4^- + OH^- \rightarrow HPO_4^{2-} + H_2O$$

and when hydronium ion is added

$$HPO_4^{2-} + H_3O^+ \rightarrow H_2PO_4^- + H_2O$$

It is apparent that the mechanism of action of this type of buffer is essentially the same as that of the weak-acid–conjugate-base buffer composed of acetic acid and sodium acetate.

Calculations—A buffer system composed of a conjugate acid–base pair, NaA – HA (such as sodium acetate and acetic acid), would have a PBE of

$$[H_3O^+] + [HA] = [OH^-] + [A^-] \tag{78}$$

Replacing [HA] and $[A^-]$ as a function of hydronium-ion concentration gives

$$[H_3O^+] + \frac{[H_3O^+]C_b}{[H_3O^+] + K_a} = [OH^-] + \frac{K_aC_a}{[H_3O^+] + K_a} \tag{79}$$

where C_b is the concentration of the salt, NaA, and C_a is the concentration of the weak acid, HA. This equation can be rearranged to give

$$[H_3O^+] = K_a \frac{(C_a - [H_3O^+] + [OH^-])}{(C_b + [H_3O^+] - [OH^-])} \tag{80}$$

In general, both C_a and C_b are much greater than $[H_3O^+]$, which is in turn much greater than $[OH^-]$, and the equation simplifies to

$$[H_3O^+] = \frac{K_aC_a}{C_b} \tag{81}$$

or, expressed in terms of pH, as

$$pH = pK_a + \log \frac{C_b}{C_a} \tag{82}$$

This equation is generally called the Henderson–Hasselbalch equation. It applies to all buffer systems formed from a single conjugate acid–base pair, regardless of the nature of the salts. For example, it applies equally well to the following buffer systems: ammonia–ammonium chloride, monosodium phosphate–disodium phosphate, phenobarbital–sodium phenobarbital, etc. In the ammonia–ammonium chloride system, ammonia is obviously the base and the ammonium ion is the acid (C_a equal to the concentration of the salt). In the phosphate system, monosodium phosphate is the acid and disodium phosphate is the base. For the phenobarbital buffer system, phenobarbital is the acid and the phenobarbital anion is the base (C_b equal to the concentration of sodium phenobarbital).

As an example of the application of this equation, the pH of a buffer solution containing acetic acid and sodium acetate, each in 0.1 *M* concentration, may be calculated. The K_a of acetic acid, as defined above, is 1.8×10^{-5}, at 25°.

Solution

First, the pK_a of acetic acid is calculated

$$pK_a = -\log K_a = -\log 1.8 \times 10^{-5}$$

$$= -\log 1.8 - \log 10^{-5}$$

$$= -0.26 - (-5) = +4.74$$

Substituting this value into Eq 82

$$pH = \log \frac{0.1}{0.1} + 4.74 = +4.74$$

The Henderson–Hasselbalch equation predicts that any solutions containing the same molar concentration of acetic acid as of sodium acetate will have the same pH. Thus, a solution of 0.01 *M* concentration of each will have the same pH, 4.74, as one of 0.1 *M* concentration of each component. Actually, there will be some difference in the pH of the solutions, for the *activity coefficient* of the components varies with concentration. For most practical purposes, however, the approximate values of pH calculated by the equation are satisfactory. It should be pointed out, however, that the buffer of higher concentration of each component will have a much greater capacity for neutralizing added acid or base and this point will be discussed further under *Buffer Capacity*.

The Henderson–Hasselbalch equation is useful also for calculating the ratio of molar concentrations of a buffer system required to produce a solution of specific pH. As an example, suppose that an acetic acid–sodium acetate buffer of pH 4.5 is to be prepared. What ratio of the buffer components should be used?

Solution

Rearranging Eq 82, which is used to calculate the pH of weak acid-salt type buffers, gives

$$\log \frac{[\text{base}]}{[\text{acid}]} = pH - pK_a$$

$$= 4.5 - 4.76 = -0.24 = (9.76 - 10)$$

$$\frac{[\text{base}]}{[\text{acid}]} = \text{antilog of } (9.76 - 10) = 0.575$$

The interpretation of this result is that the *proportion* of sodium acetate to acetic acid should be 0.575 mole of the former to 1 mole of the latter to produce a pH of 4.5. A solution containing 0.0575 mole of sodium acetate and 0.1 mole of acetic acid per liter would meet this requirement, as would also one containing 0.00575 mole of sodium acetate and 0.01 mole of acetic acid per liter. The actual concentration selected would depend chiefly on the desired buffer capacity.

Buffer Capacity—The ability of a buffer solution to resist changes in pH upon addition of acid or alkali may be measured in terms of *buffer capacity*. In the preceding discussion of buffers, it has been seen that, in a general way, the concentration of acid in a weak-acid–conjugate-base buffer determines the capacity to "neutralize" added base,

while the concentration of salt of the weak acid determines the capacity to neutralize added acid. Similarly, in a weak-base–conjugate-acid buffer the concentration of the weak base establishes the buffer capacity toward added acid, while the concentration of the conjugate acid of the weak base determines the capacity toward added base. When the buffer is equimolar in the concentrations of weak acid and conjugate base, or of weak base and conjugate acid, it has equal buffer capacity toward added strong acid or strong base.

Van Slyke, the biochemist, introduced a quantitative expression for evaluating buffer capacity. This may be defined as the amount, in gram-equivalents (g-Eq) per liter, of strong acid or strong base, required to be added to a solution to change its pH by 1 unit; a solution has a buffer capacity of 1 when 1 L requires 1 g-Eq of strong base or acid to change the pH 1 unit (in practice, considerably smaller increments are measured, expressed as the ratio of acid or base added to the change of pH produced). From this definition it is apparent that the smaller the pH change in a solution caused by the addition of a specified quantity of acid or alkali, the greater the buffer capacity of the solution.

The following numerical examples illustrate certain basic principles and calculations concerning buffer action and buffer capacity.

Example 1—What is the change of pH on adding 0.01 mole of NaOH to 1 L of 0.10 *M* acetic acid?

(a) Calculate the pH of a 0.10 molar solution of acetic acid

$$[H_3O^+] = \sqrt{K_a C_a} = \sqrt{1.75 \times 10^{-4} \times 1.0 \times 10^{-1}} = 1.33 \times 10^{-3}$$

$$pH = -\log 1.33 \times 10^{-3} = 2.88$$

(b) On adding 0.01 mole of NaOH to a liter of this solution, 0.01 mole of acetic acid is converted to 0.01 mole of sodium acetate, thereby decreasing C_a to 0.09 *M*, and $C_b = 1.0 \times 10^{-2}$ *M*. Using the Henderson–Hasselbach equation gives

$$pH = 4.76 + \log \frac{0.01}{0.09} = 4.76 - 0.95 = 3.81$$

The pH change is, therefore, 0.93 unit. The buffer capacity as defined above is calculated to be

$$\frac{\text{moles of NaOH added}}{\text{change in pH}} = 0.011$$

Example 2—What is the change of pH on adding 0.1 mole of NaOH to 1 L of buffer solution 0.1 *M* in acetic acid and 0.1 *M* in sodium acetate?

(a) The pH of the buffer solution before adding NaOH is

$$pH = \log \frac{[\text{base}]}{[\text{acid}]} + pK_a$$

$$= \log \frac{0.1}{0.1} + 4.76 = 4.76$$

(b) On adding 0.01 mole of NaOH per liter to this buffer solution, 0.01 mole of acetic acid is converted to 0.01 mole of sodium acetate, thereby decreasing the concentration of acid to 0.09 *M* and increasing the concentration of base to 0.11 *M*. The pH is calculated as

$$pH = \log \frac{0.11}{0.09} + 4.76$$

$$= 0.086 + 4.76 = 4.85$$

The change of pH in this case is only 0.09 unit, about 1/10 the change in the preceding example. The buffer capacity is calculated as

$$\frac{\text{moles of NaOH added}}{\text{change of pH}} = \frac{0.01}{0.09} = 0.11$$

Thus, the buffer capacity of the acetic acid–sodium acetate buffer solution is approximately 10 times that of the acetic acid solution.

As is in part evident from these examples, and may be further evidenced by calculations of pH changes in other systems, the degree of buffer action and, therefore, the buffer capacity, depend on the kind and concentration of the buffer components, the pH region involved and the kind of acid or alkali added.

Strong Acids and Bases as "Buffers"—In the foregoing discussion, buffer action was attributed to systems of (1) weak acids and their conjugate bases, (2) weak bases and their conjugate acids and (3) certain acid-base pairs which can function in the manner either of System 1 or 2.

The ability to resist change in pH on adding acid or alkali is possessed also by relatively concentrated solutions of strong acids and strong bases. If to 1 L of pure water having a pH of 7.0 is added 1 mL of 0.01 *M* hydrochloric acid, the pH is reduced to about 5.0. If the same volume of the acid is added to 1 L of 0.001 *M* hydrochloric acid, which has a pH of about 3, the hydronium-ion concentration is increased only about 1% and the pH is reduced hardly at all. The nature of this buffer action is quite different from that of the true buffer solutions. The very simple explanation is that when 1 mL of 0.01 *M* HCl, which represents 0.00001 g-Eq of hydronium ions, is added to the 0.0000001 g-Eq of hydronium ions in 1 L of pure water, the hydronium-ion concentration is increased 100-fold (equivalent to 2 pH units), but when the same amount is added to the 0.001 g-Eq of hydronium ions in 1 L of 0.001 *M* HCl, the increase is only 1/100 the concentration already present. Similarly, if 1 mL of 0.01 *M* NaOH is added to 1 L of pure water, the pH is increased to 9, while if the same volume is added to 1 L of 0.001 molar NaOH, the pH is increased almost immeasurably.

In general, solutions of strong acids of pH 3 or less, and solutions of strong bases of pH 11 or more, exhibit this kind of buffer action by virtue of the relatively high concentration of hydronium or hydroxyl ions present. The USP includes among its *Standard Buffer Solutions* a series of hydrochloric acid buffers, covering the pH range 1.2 to 2.2, which also contain potassium chloride. The salt does not participate in the buffering mechanism, as is the case with salts of weak acids; instead, it serves as a nonreactive constituent required to maintain the proper electrolyte environment of the solutions.

Determination of pH

Colorimetry

A relatively simple and inexpensive method for determining the approximate pH of a solution depends on the fact that some conjugate acid base pairs (indicators) possess one color in the acid form and another color in the base form. Assume that the acid form of a particular indicator is red, while the base form is yellow. The color of a solution of this indicator will range from red, when it is sufficiently acid, to yellow, when it is sufficiently alkaline. In the intermediate pH range (the transition interval) the color will be a blend of red and yellow depending upon the ratio of the base to the acid form. In general, although there are slight differences between indicators, color changes apparent to the eye cannot be discerned when the ratio of base to acid form, or acid to base form exceeds 10:1. The use of Eq 82 indicates that the transition range of most indicators is equal to the pK_a of the indicator ±1 pH unit, or a useful range of approximately 2 pH units. Standard indicator solutions can be made at known pH values within the transition range of the indicator, and the pH of an unknown solution determined by adding the indicator to it and comparing the resulting color with the standard solutions. Details of this procedure can be found in RPS-14. Another method for using these indicators is to apply them to thin strips of filter paper. A drop of the unknown solution is placed on a piece of the indicator paper and the resulting color compared to a color chart supplied with the indicator paper. These papers are available in a wide variety of pH ranges.

Potentiometry

Electrometric methods for the determination of pH are based on the fact that the difference of electrical potential between two suitable electrodes dipping into a solution containing hydronium ions depends on the concentration (or activity) of the latter. The development of a potential difference is not a specific property of hydronium ions. A solution of any ion will develop a potential proportional to the concentration of that ion if a suitable pair of electrodes is placed in the solution.

The relationship between the potential difference and concentration of an ion in equilibrium with the electrodes may be derived as follows. When a metal is immersed into a solution of one of its salts, there is a tendency for the metal to go into solution in the form of ions. This tendency is known as the *solution pressure* of the metal and is comparable to the tendency of sugar molecules, for example, to dissolve in water. The metallic ions in solution tend, on the other hand, to become discharged by forming atoms, this effect being proportional to the *osmotic pressure* of the ions. In order for an atom of a metal to go into solution as a positive ion, electrons, equal in number to the charge on the ion, must be left behind on the metal electrode with the result that the latter becomes negatively charged. The positively charged ions in solution, however, may become discharged as atoms by taking up electrons from the metal electrode. Depending on which effect predominates, the electrical charge on the electrode will be either positive or negative and may be quantitatively expressed by the following equation proposed by Nernst in 1889

$$E = \frac{RT}{nF} \ln \frac{p}{P} \tag{83}$$

where E is the potential difference or electromotive force, R is the gas constant (8.316 joules), T is the absolute temperature, n is the valence of the ion, F is the Faraday of electricity (96,500 coulombs), p is the osmotic pressure of the ions and P is the solution pressure of the metal.

Inasmuch as it is impossible to measure the potential difference between one electrode and a solution with any degree of certainty, it is customary to use two electrodes and to measure the potential difference between them. If two electrodes, both of the same metal, are immersed separately in solutions containing ions of that metal, at osmotic pressure p_1 and p_2, respectively, and connected by means of a tube containing a nonreacting salt solution (a so-called "salt-bridge"), the potential developed across the two electrodes will be equal to the difference between the potential differences of the individual electrodes; thus

$$E = E_1 - E_2 = \frac{RT}{nF} \ln \frac{p_1}{P_1} - \frac{RT}{nF} \ln \frac{p_2}{P_2} \tag{84}$$

Since both electrodes are of the same metal, $P_1 = P_2$ and the equation may be simplified to

$$E = \frac{RT}{nF} \ln p_1 - \frac{RT}{nF} \ln p_2 = \frac{RT}{nF} \ln \frac{p_1}{p_2} \tag{85}$$

In place of osmotic pressures it is permissible, for dilute solutions, to substitute the concentrations c_1 and c_2 which were found (see Chapter 16, page 222) to be proportional to p_1 and p_2. The equation then becomes

$$E = \frac{RT}{nF} \ln \frac{c_1}{c_2} \tag{86}$$

If either c_1 or c_2 is known, it is obvious that the value of the other may be found if the potential difference, E, of this cell can be measured.

For the determination of hydronium-ion concentration or pH, an electrode at which an equilibrium between hydrogen gas and hydronium ion can be established must be used in place of metallic electrodes. Such an electrode may be made by electrolytically coating a strip of platinum, or other noble metal, with platinum black and saturating the latter with pure hydrogen gas. This device functions as a *hydrogen electrode*. Two such electrodes may be assembled as shown in Fig 17-3.

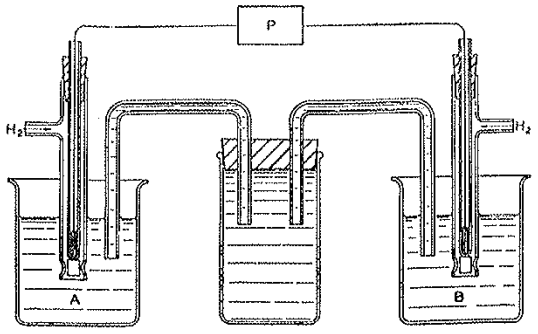


Fig 17-3. Hydrogen-ion concentration chain.

In this diagram one electrode dips into Solution A, containing a known hydronium-ion concentration, and the other electrode dips into Solution B, containing an unknown hydronium-ion concentration. The two electrodes and solutions, sometimes called half-cells, then are connected by a bridge of neutral salt solution, which has no significant effect on the solutions it connects. The potential difference across the two electrodes is measured by means of a potentiometer, P. If the concentration, c_1, of hydronium ion in Solution A is 1 N, Eq 86 simplifies to

$$E = \frac{RT}{nF} \ln \frac{1}{c_2} \tag{87}$$

or in terms of Briggsian logarithms

$$E = 2.303 \frac{RT}{nF} \log_{10} \frac{1}{c_2} \tag{88}$$

If for $\log_{10} 1/c_2$ there is substituted its equivalent pH, the equation becomes

$$E = 2.303 \frac{RT}{nF} \text{ pH} \tag{89}$$

and finally by substituting numerical values for R, n, T and F, and assuming the temperature to be 20°, the following simple relationship is derived

$$E = 0.0581 \text{ pH or pH} = \frac{E}{0.0581} \tag{90}$$

The hydrogen electrode dipping into a solution of known hydronium-ion concentration, called the *reference electrode*, may be replaced by a calomel electrode, one type of which is shown in Fig 17-4. The elements of a calomel electrode are mercury and calomel in an aqueous solution of potassium chloride. The potential of this electrode is constant, regardless of the hydronium-ion concentration of the solution into which it dips. The potential depends on the equilibrium which is set up between mercury and mercurous ions from the calomel, but the concentration of the latter is governed, according to the solubility-product principle, by the concentration of chloride ions, which are derived mainly from the potassium chloride in the solution. Therefore, the potential of this electrode varies with the concentration of potassium chloride in the electrolyte.

Because the calomel electrode always indicates voltages which are higher, by a constant value, than those obtained

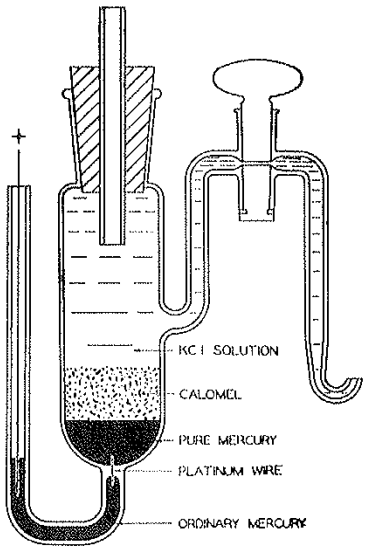


Fig 17-4. Calomel electrode.

when the normal hydrogen electrode chain shown in Fig 17-3 is used, it is necessary to subtract the potential due to the calomel electrode itself from the observed voltage. As the magnitude of this voltage depends on the concentration of potassium chloride in the calomel-electrode electrolyte, it is necessary to know the concentration of the former. For most purposes a saturated potassium chloride solution is used which produces potential difference of 0.2488 v. Accordingly, before using Eq 85 for the calculation of pH from the voltage of a cell made up of a calomel and a hydrogen electrode dipping into the solution to be tested, 0.2488 v must be subtracted from the observed potential difference. Expressed mathematically, Eq 91 is used for calculating pH from the potential difference of such a cell.

$$\text{pH} = \frac{E - 0.2488}{0.0581} \tag{91}$$

In measuring the potential difference between the electrodes, it is imperative that very little current be drawn from the cell, for with current flowing the voltage changes, owing to polarization effects at the electrode. Because of this it is not possible to make accurate measurements with a voltmeter which requires appreciable current to operate it. In its place a potentiometer is used which does not draw a current from the cell being measured.

There are many limitations to the use of the hydrogen electrode:

It cannot be used in solutions containing strong oxidants such as ferric iron, dichromates, nitric acid, peroxide or chlorine or reductants, such as sulfurous acid and hydrogen sulfide.

It is affected by the presence of organic compounds which are fairly easily reduced.

It cannot be used successfully in solutions containing cations that fall below hydrogen in the electrochemical series.

Erratic results are obtained in the measurement of unbuffered solutions unless special precautions are taken.

It is troublesome to prepare and maintain.

Since other electrodes more convenient to use now are available, the hydrogen electrode today is used rarely. Nevertheless, it is the ultimate standard for pH measurements.

To avoid some of the difficulties with the hydrogen electrode, the *quinhydrone* electrode was introduced and was popular for a long time, particularly for measurements of acid solutions. The unusual feature of this electrode is that it consists of a piece of gold or platinum wire or foil dipping into the solution to be tested, in which has been dissolved a small quantity of quinhydrone. A calomel electrode may be used for reference, just as in determinations with the hydrogen electrode.

Quinhydrone consists of an equimolecular mixture of quinone and hydroquinone; the relationship between these substances and hydrogen-ion concentration is

$$\text{Quinone} + 2\text{ hydrogen ions} + 2\text{ electrons} \rightleftharpoons \text{hydroquinone}$$

In a solution containing hydrogen ions the potential of the quinhydrone electrode is related logarithmically to hydronium-ion concentration if the ratio of the hydroquinone concentration to that of quinone is constant and practically equal to one. This ratio is maintained in an acid solution containing an excess of quinhydrone, and measurements may be made quickly and accurately; however, quinhydrone cannot be used in solutions more alkaline than pH 8.

An electrode which, because of its simplicity of operation and freedom from contamination or change of the solution being tested, has replaced both the hydrogen and quinhydrone electrodes is the *glass electrode*. It functions by virtue of the fact that when a thin membrane of a special composition of glass separates two solutions of different pH there is developed across the membrane a potential difference which depends on the pH of both solutions. If the pH of one of the solutions is known, the other may be calculated from the potential difference. In practice, the glass electrode usually consists of a bulb of the special glass fused to the end of a tube of ordinary glass. Inside the bulb is placed a solution of known pH, in contact with an internal silver-silver chloride or other electrode. This glass electrode and another reference electrode are immersed in the solution to be tested and the potential difference is measured. A potentiometer providing electronic amplification of the small current produced is employed. The modern instruments available permit reading the pH directly and provide also for compensation of variations due to temperature in the range of 0–50° and to the small but variable asymmetry potential inherent in the glass electrode.

Pharmaceutical Significance

In the broad realm of knowledge concerning the preparation and action of drugs few, if any, variables are so important as pH. For the purpose of this presentation, four principal types of pH-dependence of drug systems will be discussed: solubility, stability, activity and absorption.

Drug Solubility

If a salt, NaA, is added to water to give a concentration C_s, the following reactions occur

$$\text{Na}^+\text{A}^- \xrightarrow{\text{H}_2\text{O}} \text{Na}^+ + \text{A}^-$$

$$\text{A}^- + \text{H}_2\text{O} \rightleftharpoons \text{HA} + \text{OH}^-$$

If the pH of the solution is lowered, more of the A^- would be converted to the unionized acid, HA, in accordance with Le Chatelier's principle. Eventually, a pH will be obtained, below which the amount of HA formed exceeds its aqueous solubility, S_0, and the acid will precipitate from solution; this pH can be designated as pH_p. At this point, at which the amount of HA formed just equals S_0, a mass balance on the total amount of drug in solution yields

$$C_s = [\text{HA}] + [\text{A}^-] = S_0 + [\text{A}^-] \tag{92}$$

Replacing $[A^-]$ as a function of hydronium-ion concentration gives

$$C_s = S_0 + \frac{K_a C_s}{[H_3O^+]_p + K_a} \quad (93)$$

where K_a is the ionization constant for the conjugate acid, HA, and $[H_3O^+]_p$ refers to the hydronium-ion concentration above which precipitation will occur. This equation can be rearranged to give

$$[H_3O^+]_p = K_a \frac{S_0}{C_s - S_0} \quad (94)$$

Taking logarithms gives

$$pH_p = pK_a + \log \frac{C_s - S_0}{S_0} \quad (95)$$

Thus, the pH below which precipitation occurs is a function of the amount of salt added initially, the pK_a and the solubility of the free acid formed from the salt.

The analogous equation for salts of weak bases and strong acids (such as pilocarpine hydrochloride, cocaine hydrochloride or codeine phosphate) would be

$$pH_p = pK_a + \log \frac{S_0}{C_s - S_0} \quad (96)$$

in which pK_a refers to the protonated form of the weak base.

Example—Below what pH will free phenobarbital begin to precipitate from a solution initially containing 1.3 g of sodium phenobarbital/100 mL at 25°? The molar solubility of phenobarbital is 0.0050 and its pK_a is 7.41. The molecular weight of sodium phenobarbital is 254.

The molar concentration of salt initially added is

$$C_s = \frac{g/L}{\text{mol wt}} = \frac{13}{254} = 0.051\ M$$

$$pH_p = 7.41 + \log \frac{0.051 - 0.005}{0.005}$$

$$= 7.41 + 0.96 = 8.37$$

Example—Above what pH will free cocaine begin to precipitate from a solution initially containing 0.0294 mole of cocaine hydrochloride/L? The pK_b of cocaine is 5.59, and its molar solubility is 5.60×10^{-3}.

$$pK_a = pK_w - pK_b = 14.00 - 5.59 = 8.41$$

$$pH_p = 8.41 + \log \frac{0.0056}{0.0294 - 0.0056}$$

$$= 8.41 + (-0.63) = 7.78$$

Drug Stability

One of the most diversified and fruitful areas of study is the investigation of the effect of hydrogen-ion concentration on the stability or, in more general terms, the reactivity of pharmaceutical systems. The evidence for enhanced stability of systems when these are maintained within a narrow range of pH, as well as of progressively decreasing stability as the pH departs from the optimum range, is abundant. Stability (or instability) of a system may result from gain or loss of a proton (hydrogen ion) by a substrate molecule—often accompanied by an electronic rearrangement—which reduces (or increases) the reactivity of the molecule. Instability results when the substance desired to remain unchanged is converted to one or more other, unwanted, substances. In aqueous solution, instability may arise through the catalytic effect of acids or bases, the former by transferring a proton to the substrate molecule, the latter by accepting a proton.

Specific illustrations of the effect of hydrogen-ion concentration on the stability of medicinals are myriad; only a few will be given here, these being chosen to show the importance of pH adjustment of solutions that require sterilization.

Morphine solutions are not decomposed during a 60-min exposure at a temperature of 100° if the pH is less than 5.5; neutral and alkaline solutions, however, are highly unstable. Minimum hydrolytic decomposition of solutions of cocaine occurs in the range of pH of 2 to 5; in one study a solution of cocaine hydrochloride, initially at a pH of 5.7, remained stable during 2 months (although the pH dropped to 4.2 in this time), while another solution buffered to about pH 6 underwent approximately 30% hydrolysis in the same time. Similarly, solutions of procaine hydrochloride containing some hydrochloric acid showed no appreciable decomposition; when dissolved in water alone, 5% of the procaine hydrochloride hydrolyzed, while when buffered to pH 6.5, from 19 to 35% underwent decomposition by hydrolysis. Solutions of thiamine hydrochloride may be sterilized by autoclaving without appreciable decomposition if the pH is below 5; above this, thiamine hydrochloride is unstable.

The stability of many disperse systems, and especially of certain emulsions, is often pH-dependent. Information concerning specific emulsion systems, and the effect of pH upon them, may be found in Chapter 19.

Drug Activity

Drugs that are weak acids or weak bases, and hence may exist in ionized or nonionized form (or a mixture of both), may be active in one form but not in the other; often such drugs have an optimum pH range for maximum activity. Thus, mandelic acid, benzoic acid or salicylic acid have pronounced antibacterial activity in nonionized form but have practically no such activity in ionized form. Accordingly, these substances require an acid environment to function effectively as antibacterial agents. For example, sodium benzoate is effective as a preservative in 4% concentration at pH 7.0, in 0.06 to 0.1% concentration at pH 3.5 to 4.0 and in 0.02 to 0.03% concentration at pH 2.3 to 2.4. Other antibacterial agents, on the other hand, are active principally, if not entirely, in cationic form. Included in this category are the acridines and quaternary ammonium compounds.

Drug Absorption

The degree of ionization and lipoid solubility of a drug are two important factors that determine the rate of absorption of drugs from the gastrointestinal tract and, indeed, their passage through cellular membranes generally. Drugs which are weak organic acids or bases, and which in nonionized form are soluble in lipids, apparently are absorbed through cellular membranes by virtue of the lipoidal nature of the membranes. Completely ionized drugs, on the other hand, are absorbed poorly, if at all. Rates of absorption of a variety of drugs are related to their ionization constants and in many cases may be predicted quantitatively on the basis of this relationship. Thus, not only the degree of the acidic or basic character of a drug but consequently also the pH of the physiological medium (gastric or intestinal fluid, plasma, cerebrospinal fluid, etc) in which a drug is dissolved or dispersed—since this pH determines the extent to which the drug will be converted to ionic or nonionic form—become important parameters of drug absorption. Further information on drug absorption is given in Chapter 35.

References

1. Benet LZ, Goyan JE: *J Pharm Sci 54:* 1179, 1165.
2. Riegelman S *et al: Ibid 51:* 129, 1962.
3. Niebergall PJ *et al: Ibid 61:* 232, 1972.

Bibliography

Freiser H, Fernando Q: *Ionic Equilibria in Analytical Chemistry,* Wiley, New York, 1966.

CHAPTER 19

Disperse Systems

George Zografi, PhD
Professor
School of Pharmacy, University of Wisconsin
Madison, WI 53706

Hans Schott, PhD
Professor of Pharmaceutics and Colloid Chemistry
School of Pharmacy, Temple University
Philadelphia, PA 19140

James Swarbrick, DSc, PhD
Professor and Chairman
Division of Pharmaceutics
School of Pharmacy, University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7360

Interfacial Phenomena

Very often it is desirable or necessary in the development of pharmaceutical dosage forms to produce multiphasic dispersions by mixing together two or more ingredients which are not mutually miscible and capable of forming homogeneous solutions. Examples of such dispersions include suspensions (solid in liquid), emulsions (liquid in liquid) and foams (vapor in liquids). Because these systems are not homogeneous and thermodynamically stable, over time they will show some tendency to separate on standing to produce the minimum possible surface area of contact between phases. Thus, suspended particles agglomerate and sediment, emulsified droplets cream and coalesce and the bubbles dispersed in foams collapse, to produce unstable and nonuniform dosage forms. In this chapter the fundamental physical chemical properties of dispersed systems will be discussed, along with the principles of interfacial and colloidal physics and chemistry which underly these properties.

Interfacial Forces and Energetics

In the bulk portion of each phase, molecules are attracted to each other equally in all directions, such that no resultant forces are acting on any one molecule. The strength of these forces determines whether a substance exists as a vapor, liquid or solid at a particular temperature and pressure.

At the boundary between phases, however, molecules are acted upon unequally since they are in contact with other molecules exhibiting different forces of attraction. For example, the primary intermolecular forces in water are due to hydrogen bonds, whereas those responsible for intermolecular bonding in hydrocarbon liquids, such as mineral oil, are due to London dispersion forces.

Because of this, molecules situated at the interface contain potential forces of interaction which are not satisfied relative to the situation in each bulk phase. In liquid systems such unbalanced forces can be satisfied by spontaneous movement of molecules from the interface into the bulk phase. This leaves fewer molecules per unit area at the interface (greater intermolecular distance) and reduces the actual contact area between dissimilar molecules.

Any attempt to reverse this process by increasing the area of contact between phases, ie, bringing more molecules into the interface, causes the interface to resist expansion and to behave as though it is under a tension everywhere in a tangential direction. The force of this tension per unit length of interface generally is called the interfacial tension, except when dealing with the air–liquid interface, where the terms surface and surface tension are used.

To illustrate the presence of a tension in the interface, consider an experiment where a circular metal frame, with a looped piece of thread loosely tied to it, is dipped into a liquid. When removed and exposed to the air, a film of liquid will be stretched entirely across the circular frame, as when one uses such a frame to blow soap bubbles. Under these conditions (Fig 19-1*A*), the thread will remain collapsed. If now a heated needle is used to puncture and remove the liquid film from within the loop (Fig 19-1*B*), the loop will stretch spontaneously into a circular shape.

The result of this experiment demonstrates the spontaneous reduction of interfacial contact between air and the liquid remaining and, indeed, that a tension causing the loop to remain extended exists parallel to the interface. The circular shape of the loop indicates that the tension in the plane of the interface exists at right angles or normal to every part of the looped thread. The total force on the entire loop divided by the circumference of the circle, therefore, represents the tension per unit distance of surface, or the surface tension.

Just as work is required to extend a spring under tension, work should be required to reverse the process seen in Figs 19-1*A* and *B*, thus bringing more molecules to the interface. This may be seen quantitatively by considering an experiment where tension and work may be measured directly. Assume that we have a rectangular wire with one movable side (Fig 19-2). Assume further that by dipping this wire into a liquid, a film of liquid will form within the frame when it is removed and exposed to the air. As seen earlier in Fig 19-1, since it comes in contact with air, the liquid surface will tend to contract with a force, F, as molecules leave the surface for the bulk. To keep the movable side in equilibrium, an equal force must be applied to oppose this tension in the surface. We then may define the surface tension, γ, of the liquid as $F/2l$, where $2l$ is the distance of surface over which F is operating ($2l$ since there are two surfaces, top and bottom). If the surface is expanded by a very small distance, Δx, one can then estimate that the work done is

$$W = F\Delta x \quad (1)$$

and therefore

$$W = \gamma 2l\Delta x \quad (2)$$

Dr Zografi authored the section on *Interfacial Phenomena*. *Dr Schott* authored the section on Colloidal Dispersions. *Dr Swarbrick* authored the section on Particle Phenomena and Coarse Dispersions.

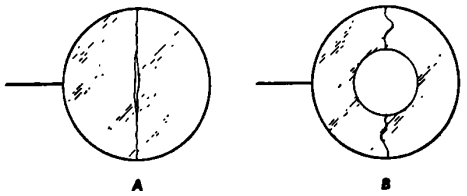


Fig 19-1. A circular wire frame with a loop of thread loosely tied to it: (*A*) a liquid film on the wire frame with a loop in it; (*B*) the film inside the loop is broken.[1]

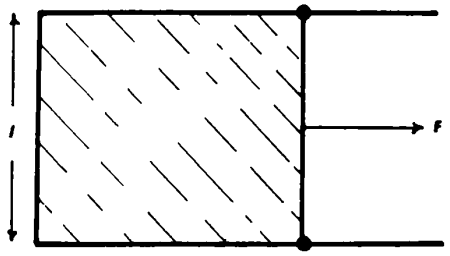


Fig 19-2. A movable wire frame containing a film of liquid being expanded with a force, *F*.

Since

$$\Delta A = 2l\Delta x \tag{3}$$

where ΔA is the change in area due to the expansion of the surface, we may conclude that

$$W = \gamma \Delta A \tag{4}$$

Thus, the work required to create a unit area of surface, known as the surface free energy/unit area, is equivalent to the surface tension of a liquid system, and the greater the area of interfacial contact between phases, the greater the free-energy increase for the total system. Since a prime requisite for equilibrium is that the free energy of a system be at a minimum, it is not surprising to observe that phases in contact tend to reduce area of contact spontaneously.

Liquids, being mobile, may assume spherical shapes (smallest interfacial area for a given volume), as when ejected from an orifice into air or when dispersed into another immiscible liquid. If a large number of drops are formed, further reduction in area can occur by having the drops coalesce, as when a foam collapses or when the liquid phases making up an emulsion separate.

Surface tension is expressed in units of dynes/cm, while surface free energy is expressed in ergs/cm^2. Since an erg is a dyne-cm, both sets of units are equivalent.

Values for the surface tension of a variety of liquids are given in Table I, while interfacial tension values for various liquids against water are given in Table II. Other combinations of immiscible phases could be given but most heterogeneous systems encountered in pharmacy usually contain water. Values for these tensions are expressed for a particular temperature. Since an increased temperature increases the thermal energy of molecules, the work required to bring molecules to the interface should be less, and thus the surface and interfacial tension will be reduced. For example, the surface tension of water at 0° is 76.5 dynes/cm and 63.5 dynes/cm at 75°.

As would be expected from the discussion so far, the relative values for surface tension should reflect the nature of intermolecular forces present; hence, the relatively large values for mercury (metallic bonds) and water (hydrogen bonds), and the lower values for benzene, chloroform, carbon tetrachloride and the *n*-alkanes. Benzene with π electrons

Table I—Surface Tension of Various Liquids at 20°

Substance	Surface tension, dynes/cm
Mercury	476
Water	72.8
Glycerin	63.4
Oleic acid	32.5
Benzene	28.9
Chloroform	27.1
Carbon tetrachloride	26.8
1-Octanol	26.5
Hexadecane	27.4
Dodecane	25.4
Decane	23.9
Octane	21.8
Heptane	19.7
Hexane	18.0
Perfluoroheptane	11.0
Nitrogen (at 75°K)	9.4

Table II—Interfacial Tension of Various Liquids against Water at 20°

Substance	Interfacial tension, dynes/cm
Decane	52.3
Octane	51.7
Hexane	50.8
Carbon tetrachloride	45.0
Chloroform	32.8
Benzene	35.0
Mercury	428
Oleic acid	15.6
1-Octanol	8.51

exhibits a higher surface tension than the alkanes of comparable molecular weight, but increasing the molecular weight of the alkanes (and hence intermolecular attraction) increases their surface tension closer to that of benzene. The lower values for the more nonpolar substances, perfluoroheptane and liquid nitrogen, demonstrate this point even more strongly.

Values of interfacial tension should reflect the differences in chemical structure of the two phases involved; the greater the tendency to interact, the less the interfacial tension. The 20-dynes/cm difference between air–water tension and that at the octane–water interface reflects the small but significant interaction between octane molecules and water molecules at the interface. This is seen also in Table II, by comparing values for octane and octanol, oleic acid and the alkanes, or chloroform and carbon tetrachloride.

In each case the presence of chemical groups capable of hydrogen bonding with water markedly reduces the interfacial tension, presumably by satisfying the unbalanced forces at the interface. These observations strongly suggest that molecules at an interface arrange themselves or orient so as to minimize differences between bulk phases.

That this occurs even at the air–liquid interface is seen when one notes the relatively low surface-tension values of very different chemical structures such as the *n*-alkanes, octanol, oleic acid, benzene and chloroform. Presumably, in each case, the similar nonpolar groups are oriented toward the air with any polar groups oriented away toward the bulk phase. This tendency for molecules to orient at an interface is a basic factor in interfacial phenomena and will be discussed more fully in succeeding sections.

Solid substances such as metals, metal oxides, silicates and salts, all containing polar groups exposed at their surface, may be classified as high-energy solids, whereas nonpo-

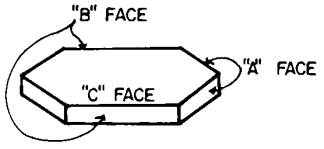


Fig 19-3. Adipic acid crystal showing various faces.[2]

Table III—Values of γ_{sv} for Solids of Varying Polarity

Solid	γ_{sv} (dynes/cm)
Teflon	19.0
Paraffin	25.5
Polyethylene	37.6
Polymethyl methacrylate	45.4
Nylon	50.8
Indomethacin	61.8
Griseofulvin	62.2
Hydrocortisone	68.7
Sodium Chloride	155
Copper	1300

lar solids such as carbon, sulfur, glyceryl tristearate, polyethylene and polytetrafluoroethylene (Teflon) may be classified as low-energy solids. It is of interest to measure the surface free energy of solids; however, the lack of mobility of molecules at the surface of solids prevents the observation and direct measurement of a surface tension. It is possible to measure the work required to create new solid surface by cleaving a crystal and measuring the work involved. However, this work not only represents free energy due to exposed groups but also takes into account the mechanical energy associated with the crystal (ie, plastic and elastic deformation and strain energies due to crystal structure and imperfections in that structure).

Also contributing to the complexity of a solid surface is the heterogeneous behavior due to the exposure of different crystal faces, each having a different surface free energy/unit area. For example, adipic acid, $HOOC(CH_2)_4COOH$, crystallizes from water as thin hexagonal plates with three different faces, as shown in Fig 19-3. Each unit cell of such a crystal contains adipic acid molecules oriented such that the hexagonal planes (faces) contain exposed carboxyl groups, while the sides and edges (A and B faces) represent the side view of the carboxyl and alkyl groups, and thus are quite nonpolar. Indeed, interactions involving these different faces reflect the differing surface free energies.[2]

Other complexities associated with solid surfaces include surface roughness, porosity and the defects and contamination produced during a recrystallization or comminution of the solid. In view of all these complications, surface free energy values for solids, when reported, should be regarded as average values, often dependent on the method used and not necessarily the same for other samples of the same substance.

In Table III are listed some approximate average values of γ_{sv} for a variety of solids, ranging in polarity from Teflon to copper, obtained by various indirect techniques.

Adhesional and Cohesional Forces

Of prime importance to those dealing with heterogeneous systems is the question of how two phases will behave when brought in contact with each other. It is well known, for instance, that some liquids, when placed in contact with other liquid or solid surfaces, will remain retracted in the form of a drop (known as a lens), while other liquids may exhibit a tendency to spread and cover the surface of this liquid or solid.

Based upon concepts developed to this point, it is apparent that the individual phases will exhibit a tendency to minimize the area of contact with other phases, thus leading to phase separation. On the other hand, the tendency for interaction between molecules at the new interface will offset this to some extent and give rise to the spontaneous spreading of one substance over the other.

In essence, therefore, phase affinity is increased as the forces of attraction between different phases (adhesional forces) become greater than the forces of attraction between molecules of the same phase (cohesional forces). If these adhesional forces become great enough, miscibility will occur and the interface will disappear. The present discussion is concerned only with systems of limited phase affinity, where an interface still exists.

A convenient approach used to express these forces quantitatively involves the use of the terms work of adhesion and work of cohesion.

The work of adhesion, W_a, is defined as the energy per cm^2 required to separate two phases at their boundary and is equal but opposite in sign to the free energy/cm^2 released when the interface is formed. In an analogous manner the work of cohesion for a pure substance, W_c, is the work/cm^2 required to produce two new surfaces, as when separating different phases, but now both surfaces contain the same molecules. This is equal and opposite in sign to the free energy/cm^2 released when the same two pure liquid surfaces are brought together and eliminated.

By convention, when the work of adhesion between two substances, A and B, exceeds the work of cohesion for one substance, eg, B, spontaneous spreading of B over the surface of A should occur with a net loss of free energy equal to the difference between W_a and W_c. If W_c exceeds W_a, no spontaneous spreading of B over A can occur. The difference between W_a and W_c is known as the spreading coefficient, S; only when S is positive will spreading occur.

The values for W_a and W_c (and hence S) may be expressed in terms of surface and interfacial tensions, when one considers that upon separation of two phases, A and B, γ_{AB} ergs of interfacial free energy/cm^2 (interfacial tension) are lost, but that γ_A and γ_B ergs/cm^2 of energy (surface tensions of A and B) are gained; upon separation of bulk phase molecules in an analogous manner, $2\gamma_A$ or $2\gamma_B$ ergs/cm^2 will be gained. Thus

$$W_a = \gamma_A + \gamma_B - \gamma_{AB} \tag{5}$$

and

$$W_c = 2\gamma_A \text{ or } 2\gamma_B \tag{6}$$

For B spreading on the surface of A, therefore

$$S_B = \gamma_A + \gamma_B - \gamma_{AB} - 2\gamma_B \tag{7}$$

or

$$S_B = \gamma_A - (\gamma_B + \gamma_{AB}) \tag{8}$$

Utilizing Eq 8 and values of surface and interfacial tension given in Tables I and II, S can be calculated for three representative substances—decane, benzene, and oleic acid—on water at 20°.

Decane:	$S = 72.8 - (23.9 + 52.3)$	$= -3.4$
Benzene:	$S = 72.8 - (28.9 + 35.0)$	$= 8.9$
Oleic acid:	$S = 72.8 - (32.5 + 15.6)$	$= 24.7$

As expected, relatively nonpolar substances such as decane exhibit negative values of S, whereas the more polar materials yield positive values; the greater the polarity of the mole-

cule, the more positive the value of S. The importance of the cohesive energy of the spreading liquid may be noted also by comparing the spreading coefficients for hexane on water and water on hexane:

$$S_{H/W} = 72.8 - (18.0 + 50.8) = \quad 4.0$$

$$S_{W/H} = 18.0 - (72.8 + 50.8) = -105.6$$

Here, despite the fact that both liquids are the same, the high cohesion and air–liquid tension of water prevents spreading on the low-energy hexane surface, while the very low value for hexane allows spreading on the water surface. This also is seen when comparing the positive spreading coefficient of hexane to the negative value for decane on water.

To see whether spreading does or does not occur, a powder such as talc or charcoal can be sprinkled over the surface of water such that it floats; then, a drop of each liquid is placed on this surface. As predicted, decane will remain as an intact drop, while hexane, benzene and oleic acid will spread out, as shown by the rapid movement of solid particles away from the point where the liquid drop was placed originally.

An apparent contradiction to these observations may be noted for hexane, benzene and oleic acid when more of each substance is added, in that lenses now appear to form even though initial spreading occurred. Thus, in effect a substance does not appear to spread over itself.

It is now established that the spreading substance forms a monomolecular film which creates a new surface having a lower surface free energy than pure water. This arises because of the apparent orientation of the molecules in such a film so that their most hydrophobic portion is oriented towards the spreading phase. It is the lack of affinity between this exposed portion of the spread molecules and the polar portion of the remaining molecules which prevents further spreading.

This may be seen by calculating a final spreading coefficient where the new surface tension of water plus monomolecular film is used. For example, the presence of benzene reduces the surface tension of water to 62.2 dynes/cm so that the final spreading coefficient, S_F, is

$$S_F = 62.2 - (28.9 + 35.0) = -1.7$$

The lack of spreading exhibited by oleic acid should be reflected in an even more negative final spreading coefficient, since the very polar carboxyl groups should have very little affinity for the exposed alkyl chain of the oleic acid film. Spreading so as to form a second layer with polar groups exposed to the air would also seem very unlikely, thus leading to the formation of a lens.

Wetting Phenomena

In the experiment described above it was shown that talc or charcoal sprinkled onto the surface of water float despite the fact that their densities are much greater than that of water. In order for immersion of the solid to occur, the liquid must displace air and spread over the surface of the solid; when liquids cannot spread over a solid surface spontaneously, and, therefore, S, the spreading coefficient, is negative, we say that the solid is not wetted.

An important parameter which reflects the degree of wetting is the angle which the liquid makes with the solid surface at the point of contact (Fig 19-4). By convention, when wetting is complete, the contact angle is zero; in nonwetting situations it theoretically can increase to a value of 180°, where a spherical droplet makes contact with solid at only one point.

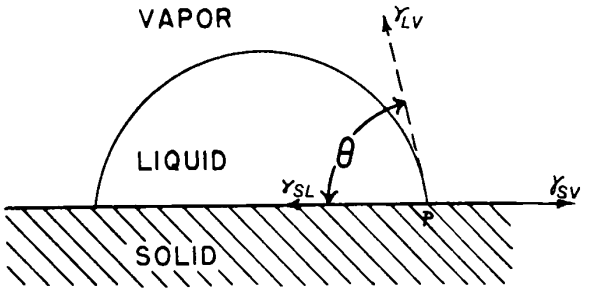


Fig 19-4. Forces acting on a nonwetting liquid drop exhibiting a contact angle of θ.[3]

In order to express contact angle in terms of solid–liquid–air equilibria, one can balance forces parallel to the solid surface at the point of contact between all three phases (Fig 19-4), as expressed in

$$\gamma_{SV} = \gamma_{SL} + \gamma_{LV} \cos \theta \qquad (9)$$

where γ_{SV}, γ_{SL}, and γ_{LV} represent the surface free energy/unit area of the solid–air, solid–liquid, and liquid–air interfaces, respectively. Although difficult to use quantitatively because of uncertainties with γ_{SV} and γ_{SL} measurements, conceptually the equation, known as the Young equation, is useful because it shows that the loss of free energy due to elimination of the air–solid interface by wetting is offset by the increased solid–liquid and liquid–air area of contact as the drop spreads out.

The $\gamma_{LV} \cos \theta$ term arises as the horizontal vectorial component of the force acting along the surface of the drop, as represented by γ_{LV}. Factors tending to reduce γ_{LV} and γ_{SL}, therefore, will favor wetting, while the greater the value of γ_{SV} the greater the chance for wetting to occur. This is seen in Table IV for the wetting of a low-energy surface, paraffin (hydrocarbon), and a higher energy surface, nylon, (polyhexamethylene adipamide). Here, the lower the surface tension of a liquid, the smaller the contact angle on a given solid, and the more polar the solid, the smaller the contact angle with the same liquid.

With Eq 9 in mind and looking at Fig 19-5, it is now possible to understand how the forces acting at the solid-

Table IV—Contact Angle on Paraffin and Nylon for Various Liquids of Differing Surface Tension

Substance	Surface tension, dynes/cm	Contact angle Paraffin	Nylon
Water	72.8	105°	70°
Glycerin	63.4	96°	60°
Formamide	58.2	91°	50°
Methylene iodide	50.8	66°	41°
α-Bromonaphthalene	44.6	47°	16°
tert-Butylnaphthalene	33.7	38°	spreads
Benzene	28.9	24°	"
Dodecane	25.4	17°	"
Decane	23.9	7°	"
Nonane	22.9	spreads	"

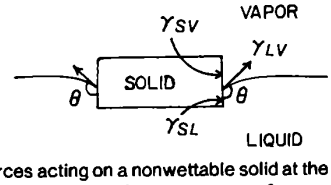


Fig 19-5. Forces acting on a nonwettable solid at the air+liquid+solid interface: contact angle θ greater than 90°.

Table V—Critical Surface Tensions of Various Polymeric Solids

Polymeric Solid	γ_c, Dynes/cm at 20°
Polymethacrylic ester of ϕ'-octanol	10.6
Polyhexafluoropropylene	16.2
Polytetrafluoroethylene	18.5
Polytrifluoroethylene	22
Poly(vinylidene fluoride)	25
Poly(vinyl fluoride)	28
Polyethylene	31
Polytrifluorochloroethylene	31
Polystyrene	33
Poly(vinyl alcohol)	37
Poly(methyl methacrylate)	39
Poly(vinyl chloride)	39
Poly(vinylidene chloride)	40
Poly(ethylene terephthalate)	43
Poly(hexamethylene adipamide)	46

liquid-air interface can cause a dense nonwetted solid to float if γ_{SL} and γ_{LV} are large enough relative to γ_{SV}.

The significance of reducing γ_{LV} was first developed empirically by Zisman when he plotted cos θ vs the surface tension of a series of liquids and found that a linear relationship, dependent on the solid, was obtained. When such plots are extrapolated to cos θ equal to one or a zero contact angle, a value of surface tension required to just cause complete wetting is obtained. Doing this for a number of solids, it was shown that this surface tension (known as the critical surface tension, γ_c) parallels expected solid surface energy γ_{SV}; the lower γ_c, the more nonpolar the surface.

Table V indicates some of these γ_c values for different surface groups, indicating such a trend. Thus, water with a surface tension of about 72 dynes/cm will not wet polyethylene (γ_c = 31 dynes/cm), but heptane with a surface tension of about 20 dynes/cm will. Likewise, Teflon (polytetrafluoroethylene) (γ_c = 19) is not wetted by heptane but is wetted by perfluoroheptane with a surface tension of 11 dynes/cm.

One complication associated with the wetting of high-energy surfaces is the lack of wetting after the initial formation of a monomolecular film by the spreading substance. As in the case of oleic acid spreading on the surface of water, the remaining liquid retracts because of the low-energy surface produced by the oriented film. This phenomenon, often called autophobic behavior, is an important factor in many systems of pharmaceutical interest since many solids, expected to be wetted easily by water, may be rendered hydrophobic if other molecules dissolved in the water can form these monomolecular films at the solid surface.

Capillarity

Because water shows a strong tendency to spread out over a polar surface such as clean glass (contact angle 0°), one would expect to observe the meniscus which forms when water is contained in a glass vessel such as a pipet or buret. This behavior is accentuated dramatically if a fine-bore capillary tube is placed into the liquid (Fig 19-6); not only will the wetting of the glass produce a more highly curved meniscus, but the level of the liquid in the tube will be appreciably higher than the level of the water in the beaker.

The spontaneous movement of a liquid into a capillary or narrow tube due to surface forces is defined as capillarity and is responsible for a number of important processes involving the penetration of liquids into porous solids. In contrast to water in contact with glass, if the same capillary is placed into mercury (contact angle on glass: 130°), not

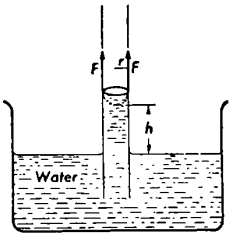


Fig 19-6. Capillary rise for a liquid exhibiting zero contact angle [1]

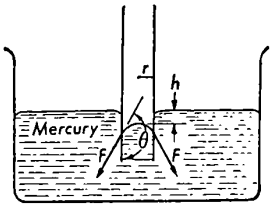


Fig 19-7. Capillary fall for a liquid exhibiting a contact angle, θ, which is greater than 90°.[1]

only will the meniscus be inverted (see Fig 19-7), but the level of the mercury in the capillary will be lower than in the beaker. In this case one does not expect mercury or other *nonwetting* liquids to easily penetrate pores unless external forces are applied.

To quantitate the factors giving rise to the phenomenon of capillarity, let us consider the case of a liquid which rises to a height, h, above the bulk liquid in a capillary having a radius, r. If (as shown in Fig 19-6) the contact angle of water on glass is zero, a force, F, will act upward and vertically along the circle of liquid–glass contact. Based upon the definition of surface tension this force will be equal to the surface tension, γ, multiplied by the circumference of the circle, $2\pi r$. Thus

$$F = \gamma 2\pi r \tag{10}$$

This force upward must support the column of water, and since the mass, m, of the column is equal to the density, d, multiplied by the volume of the column, $\pi r^2 h$, the force W opposing the movement upward will be

$$W = mg = \pi r^2 dgh \tag{11}$$

where g is the gravity constant.

Equating the two forces at equilibrium gives

$$\pi r^2 dgh = \gamma 2\pi r \tag{12}$$

so that

$$h = \frac{2\gamma}{rdg} \tag{13}$$

Thus, the greater the surface tension and the finer the capillary radius, the greater the rise of liquid in the capillary.

If the contact angle of liquid is not zero (as shown in Fig 19-8), the same relationship may be developed, except the

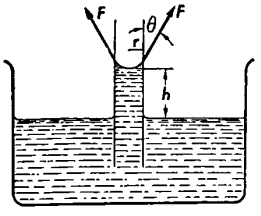


Fig 19-8. Capillary rise for a liquid exhibiting a contact angle, θ, which is greater than zero but less than 90°.[1]

vertical component of F which opposes the weight of the column is $F \cos \theta$ and, therefore

$$h = \frac{2\gamma \cos \theta}{rdg} \quad (14)$$

This indicates the very important fact that if θ is less than 90°, but greater than 0°, the value of h will decrease with increasing contact angle until at 90° ($\cos \theta = 0$), $h = 0$. Above 90°, values of h will be negative, as indicated in Fig 19-7 for mercury. Thus, based on these equations we may conclude that capillarity will occur spontaneously in a cylindrical pore even if the contact angle is greater than zero, but it will not occur at all if the contact angle becomes 90° or more. In solids with irregularly shaped pores the relationships between parameters in Eq 14 will be the same, but they will be more difficult to quantitate because of nonuniform changes in pore radius throughout the porous structure.

Pressure Differences across Curved Surfaces

From the preceding discussion of capillarity another important concept follows. In order for the liquid in a capillary to rise spontaneously it must develop a higher pressure than the lower level of the liquid in the beaker. However, since the system is open to the atmosphere, both surfaces are in equilibrium with the atmospheric pressure. In order to be raised above the level of liquid in the beaker and produce a hydrostatic pressure equal to hgd, the pressure just below the liquid meniscus, in the capillary, P_1, must be less than that just below the flat liquid surface, P_0, by hgd, and therefore

$$P_0 - P_1 = hgd \quad (15)$$

Since, according to Eq 14

$$h = \frac{2\gamma \cos \theta}{rgd}$$

then

$$P_0 - P_1 = \frac{2\gamma \cos \theta}{r} \quad (16)$$

For a contact angle of zero, where the radius of the capillary is the radius of the hemisphere making up the meniscus,

$$P_0 - P_1 = \frac{2\gamma}{r} \quad (17)$$

The consequences of this relationship (known as the Laplace equation) are important for any curved surface when r becomes very small and γ is relatively significant. For example, a spherical droplet of air formed in a bulk liquid and having a radius, r, will have a greater pressure on the inner concave surface than on the convex side, as expressed in Eq 17.

Another direct consequence of what Eq 17 expresses is the fact that very small droplets of liquid, having highly curved surfaces, will exhibit a higher vapor pressure, P, than that observed over a flat surface of the same liquid at P'. The equation (Eq. 18) expressing the ratio of P/P' to droplet radius, r, and surface tension, γ, is called the Kelvin equation where

$$\log P/P' = \frac{2\gamma M}{2.303RT\rho r} \quad (18)$$

and M is the molecular weight, R the gas constant in ergs per mole per degree, T is temperature and ρ is the density in g/cm^3. Values for the ratio of vapor pressures are given in Table VI for water droplets of varying size. Such ratios indicate why it is possible for very fine water droplets in clouds to remain uncondensed despite their close proximity to one another.

Table VI—Ratio of Observed Vapor Pressure to Expected Vapor Pressure of Water at 25° with Varying Droplet Size

P/P' [a]	Droplet size, μm
1.001	1
1.01	0.1
1.1	0.01
2.0	0.005
3.0	0.001
4.2	0.00065
5.2	0.00060

[a] P is the observed vapor pressure and P' is the expected value for "bulk" water.

This same behavior may be seen when measuring the solubility of very fine solid particles since both vapor pressure and solubility are measures of the escaping tendency of molecules from a surface. Indeed, the equilibrium solubility of extremely small particles has been shown to be greater than the usual value noted for coarser particles; the greater the surface energy and smaller the particles, the greater this effect.

Adsorption

Vapor Adsorption on Solid Surfaces

It was suggested earlier that a high surface or interfacial free energy may exist at a solid surface if the unbalanced forces at the surface and the area of exposed groups are quite great.

Substances such as metals, metal oxides, silicates, and salts—all containing exposed polar groups—may be classified as high-energy or hydrophilic solids; nonpolar solids such as carbon, sulfur, polyethylene, or Teflon (polytetrafluoroethylene) may be classified as low-energy or hydrophobic solids (Table III). Whereas liquids satisfy their unbalanced surface forces by changes in shape, pure solids (which exhibit negligible surface mobility) must rely on reaction with molecules either in the vapor state or in a solution which comes in contact with the solid surface to accomplish this.

Vapor adsorption is the simplest model demonstrating how solids reduce their surface free energy in this manner.

Depending on the chemical nature of the adsorbent (solid) and the adsorbate (vapor), the strength of interaction between the two species may vary from strong specific chemical bonding to interactions produced by the weaker more nonspecific London dispersion forces. Ordinarily, these latter forces are those responsible for the condensation of relatively nonpolar substances such as N_2, O_2, CO_2 or hydrocarbons.

When chemical reaction occurs, the process is called chemisorption; when dispersion forces predominate, the term physisorption is used. Physisorption occurs at temperatures approaching the liquefaction temperature of the vapor, whereas, for chemisorption, temperatures depend on the particular reaction involved. Water-vapor adsorption to various polar solids can occur at room temperature through hydrogen-bonding, with binding energies intermediate to physisorption and chemisorption.

In order to study the adsorption of vapors onto solid surfaces one must measure the amount of gas adsorbed/unit area or unit mass of solid, at different pressures of gas. Since such studies usually are conducted at constant temperature, plots of volume adsorbed vs pressure are referred to as adsorption isotherms. If the physical or chemical adsorption process is monomolecular, the adsorption iso-

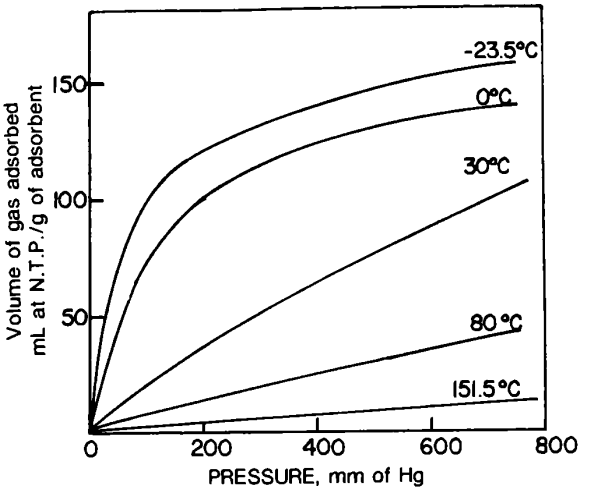


Fig 19-9 Adsorption isotherms for ammonia on charcoal.[4]

therm should look like those shown in Fig 19-9. Note the significant increase in adsorption with increasing pressure, followed by a leveling off. This leveling off is due either to a saturation of available specific chemical groups, as in chemisorption, or to the entire available surface being covered by physically adsorbed molecules. Note also the reduction in adsorption with increasing temperature which occurs because the adsorption process is exothermic. Often in the case of physical adsorption at low temperatures, after adsorption levels off, a marked increase in adsorption occurs, presumably due to multilayered adsorption. In this case vapor molecules essentially condense upon themselves as the liquefaction pressure of the vapor is approached. Figure 19-10 illustrates one type of isotherm generally seen with multilayered physisorption.

In order to have some quantitative understanding of the adsorption process and to be able to compare different systems, two factors must be evaluated; it is important to know what the capacity of the solid is or what the maximum amount of adsorption is under a given set of conditions and what the affinity of a given substance is for the solid surface or how readily does it adsorb for a given amount of pressure? In effect, this second term is the equilibrium constant for the process.

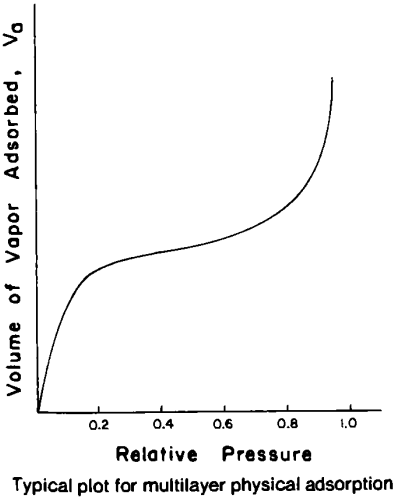


Fig 19-10. Typical plot for multilayer physical adsorption of a vapor on a solid surface.

A significant development along these lines was introduced by Langmuir when he proposed his theory of monomolecular adsorption. He postulated that for adsorption to occur a solid must contain uniform adsorption sites, each capable of holding a gas molecule. Molecules colliding with the surface may bounce off elastically or they may remain in contact for a period of time. It is this contact over a period of time that Langmuir termed adsorption.

Two major assumptions were made in deriving the equation: (1) only those molecules striking an empty site can be adsorbed, hence, only monomolecular adsorption occurs, and (2) the forces of interaction between adsorbed molecules are negligible and, therefore, the probability of a molecule adsorbing onto or desorbing from any site is independent of the surrounding sites.

The derivation of the equation is based upon the relationship between the rate of adsorption and desorption, since at equilibrium the two rates must be equal. Let μ equal the number of molecules striking each sq cm of surface/sec. From the kinetic theory of gases

$$\mu = \frac{p}{(2\pi mkT)^{1/2}} \tag{19}$$

where p is the gas pressure, m is the mass of the molecule, k is the Boltzmann gas constant, and T is the absolute temperature. Thus, the greater p, the greater the number of collisions. Let α equal the fraction of molecules which will be held by the surface; then $\alpha\mu$ is equal to the rate of adsorption on the bare surface. However, if θ is the fraction of the surface already covered, the rate of adsorption actually will be

$$R_a = \alpha\mu(1 - \theta) \tag{20}$$

In a similar manner the rate of molecules leaving the surface can be expressed as

$$R_d = \gamma\theta \tag{21}$$

where γ is the rate at which molecules can leave the surface and θ represents the number of molecules available to desorb. The value of γ strongly depends on the energy associated with adsorption; the greater the binding energy, the lower the value of γ. At equilibrium, $R_a = R_d$ and

$$\gamma\theta = \alpha\mu(1 - \theta) \tag{22}$$

Isolating the variable term, p, and combining all constants into k, the equation can be written as

$$\theta = \frac{kp}{1 + kp} \tag{23}$$

and, since θ may be expressed as

$$\theta = \frac{V_a}{V_m} \tag{24}$$

where V_a is the volume of gas adsorbed and V_m is the volume of gas covering all of the sites, Eq. 23 may be written as

$$V_a = \frac{V_m kp}{1 + kp} \tag{25}$$

A test of fit to this equation can be made by expressing it in linear form

$$\frac{p}{V_a} = \frac{1}{V_m k} + \frac{p}{V_m} \tag{26}$$

The value of k is, in effect, the equilibrium constant and may be used to compare affinities of different substances for the solid surface. The value of V_m is valuable since it indicates the maximum number of sites available for adsorption. In the case of physisorption the maximum number of sites is

actually the total surface area of the solid and, therefore, the value of V_m can be used to estimate surface area if the volume and area/molecule of vapor are known.

Since physisorption most often involves some multilayered adsorption, an equation, based on the Langmuir equation, the B.E.T. equation, is normally used to determine V_m and solid surface areas. Equation 27 is the B.E.T. equation:

$$V_a = \frac{V_m c p}{(p_0 - p)[1 + (C - 1)(p/p_0]} \qquad (27)$$

where c is a constant and p_0 is the vapor pressure of the adsorbing substance.[5] The most widely used vapor for this purpose is nitrogen, which adsorbs nonspecifically on most solids near its boiling point at −195° and appears to occupy about 16 Å^2/molecule on a solid surface.

Adsorption from Solution

By far one of the most important aspects of interfacial phenomena encountered in pharmaceutical systems is the tendency for substances dissolved in a liquid to adsorb to various interfaces. Adsorption from solution is generally more complex than that from the vapor state because of the influence of the solvent and any other solutes dissolved in the solvent. Although such adsorption is generally limited to one molecular layer, the presence of other molecules often makes the interpretation of adsorption mechanisms much more difficult than for chemisorption or physisorption of a vapor. Since monomolecular adsorption from solution is so widespread at all interfaces, we will first discuss the nature of monomolecular films and then return to a discussion of adsorption from solution.

Insoluble Monomolecular Films

It was suggested above that molecules exhibiting a tendency to spread out at an interface might be expected to orient so as to reduce the interfacial free energy produced by the presence of the interface. Direct evidence for molecular orientation has been obtained from studies dealing with the spreading on water of insoluble polar substances containing long hydrocarbon chains, eg, fatty acids.

In the late 19th century Pockels and Rayleigh showed that a very small amount of olive or castor oil—when placed on the surface of water—spreads out, as discussed above. If the amount of material was less than could physically cover the entire surface only a slight reduction in the surface tension of water was noted. However, if the surface was compressed between barriers, as shown in Fig 19-11, the surface tension was reduced considerably.

Devaux extended the use of this technique by dissolving small amounts of solid in volatile solvents and dropping the solution onto a water surface. After assisting the water-insoluble molecules to spread, the solvent evaporated, leaving a surface film containing a known amount of solute.

Compression and measurement of surface tension indicated that a maximum reduction of surface was reached when the number of molecules/unit area was reduced to a value corresponding to complete coverage of the surface. This suggested that a monomolecular film forms and that surface tension is reduced upon compression because contact between air and water is reduced by the presence of the film molecules. Beyond the point of closest packing the film apparently collapses very much as a layer of corks floating on water would be disrupted when laterally compressed beyond the point of initial physical contact.

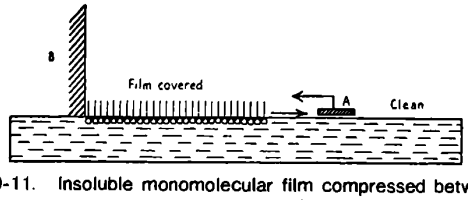


Fig 19-11. Insoluble monomolecular film compressed between a fixed barrier, *B*, and a movable barrier, *A*.[6]

Using a refined quantitative technique based on these studies, Langmuir[7] spread films of pure fatty acids, alcohols, and esters on the surface of water. Comparing a series of saturated fatty acids, differing only in chain length, he found that the area/molecule at collapse was independent of chain length, corresponding to the cross-sectional area of a molecule oriented in a vertical position (see Fig 19-11). He further concluded that this molecular orientation involved association of the polar carboxyl group with the water phase and the nonpolar acyl chain out towards the vapor phase.

In addition to the evidence for molecular orientation, Langmuir's work with surface films revealed that each substance exhibits film properties which reflect the interactions between molecules in the surface film. This is best seen by plotting the difference in surface tension of the clean surface, γ_0, and that of the surface covered with the film, γ, vs the area/molecule, A, produced by film compression (total area ÷ the number of molecules). The difference in surface tension is called the surface pressure, π, and thus

$$\pi = \gamma_0 - \gamma. \qquad (28)$$

Figure 19-12 depicts such a plot for a typical fatty acid monomolecular film. At areas greater than 50 Å^2/molecule the molecules are far apart and do not cover enough surface to reduce the surface tension of the clean surface to any extent and thus the lack of appreciable surface pressure. Since the molecules in the film are quite free to move laterally in the surface, they are said to be in a two-dimensional "gaseous" or "vapor" state.

As the intermolecular distance is reduced upon compression, the surface pressure rises because the air–water surface is being covered to a greater extent. The rate of change in π with A, however, will depend on the extent of interaction between film molecules; the greater the rate of change, the more "condensed" the state of the film.

In Fig 19-12, from 50 Å^2 to 30 Å^2/molecule, the curve shows a steady increase in π, representative of a two-dimensional "liquid" film, where the molecules become more restricted in their freedom of movement because of interactions. Below 30 Å^2/molecule the increase in π occurs over a narrow range of A, characteristic of closest packing and a two-dimensional "solid" film.

Any factor tending to increase polarity or bulkiness of the molecule—such as increased charge, number of polar groups, reduction in chain length, or the introduction of

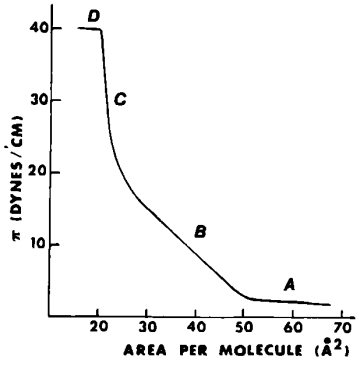


Fig 19-12. A surface pressure–area curve for an insoluble monomolecular film: Region *A*, "gaseous" film; Region *B*, "liquid" film; Region *C*, "solid" film; Region *D*, film collapse

aromatic rings, side chains, and double bonds—should reduce molecular interactions, while the longer the alkyl chain and the less bulky the polar group, the closer the molecules can approach and the stronger the extent of interaction in the film.

Soluble Films and Adsorption from Solution

If a fatty acid exhibits highly "gaseous" film behavior on an aqueous surface, we should expect a relatively small change in π with A over a considerable range of compression. Indeed, for short-chain compounds—eg, lauric acid (12 carbons) and decanoic acid—not only is the change in π small with decreasing A but at a point just before the expected closest packing area the surface pressure becomes constant without any collapse.

If lauric acid is converted to the laurate ion, or if a shorter chain acid such as octanoic acid is used, spreading on water and compression of the surface produces no increase in π; the more polar the molecule (hence, the more "gaseous" the film), the higher the area/molecule where a constant surface pressure occurs.

This behavior may be explained by assuming that polar molecules form monomolecular films when spread on water but that, upon compression, they are caused to enter the aqueous bulk solution rather than to remain as an intact insoluble film. The constant surface pressure with increased compression arises because a constant number of molecules/unit area remain at the surface in equilibrium with dissolved molecules. The extent of such behavior will be greater for substances exhibiting weaker intermolecular interaction and greater water solubility.

Starting from the other direction, it can be shown that short-chain acids and alcohols (when dissolved in water) reduce the surface tension of water, thus producing a surface pressure, just as with insoluble films (see Eq 28). That dissolved molecules are accumulating at the interface in the form of a monomolecular film is suggested from the similarity in behavior to systems where slightly soluble molecules are spread on the surface. For example, compressing the surface of a solution containing "surface-active" molecules has no effect on the initial surface pressure, whereas increasing bulk-solution concentration tends to increase surface pressure, presumably by shifting the equilibrium between surface and bulk molecules.

At this point we may ask, why should water-soluble molecules leave an aqueous phase and accumulate or "adsorb" at an air–solution interface? Since any process will occur spontaneously if it results in a net loss in free energy, such must be the case for the process of adsorption.

A number of factors will produce such a favorable change in free energy. First, the presence of the oriented monomolecular film reduces the surface free energy of the air–water interface. Second, the hydrophobic group on the molecule is in a lower state of energy at the interface, where it no longer is as surrounded by water molecules, than when it is in the bulk-solution phase. Increased interaction between film molecules also will contribute to this process.

A further reduction in free energy occurs upon adsorption because of the gain in entropy associated with a change in water structure. Water molecules, in the presence of dissolved alkyl chains are more highly organized or "ice-like" than they are as a pure bulk phase; hence, the entropy of such structured water is lower than that of bulk water.

The process of adsorption requires that the "ice-like" structure "melt" as the chains go to the interface and, thus, an increase in the entropy of water occurs. The adsorption of molecules dissolved in oil can occur but it is not influenced by water structure changes and, hence, only the first factors mentioned are important here.

It is very rare that significant adsorption can occur at the hydrocarbon–air interface since little loss in free energy can occur by bringing hydrocarbon chains with polar groups attached to this interface; however, at oil–water interfaces the polar portions of the molecule can interact with water at the interface, leading to significant adsorption.

Thus, whereas water-soluble fatty acid salts are adsorbed from water to air–water and oil–water interfaces, their undissociated counterparts, the free fatty acids, which are water insoluble, form insoluble films at the air–water interface, are not adsorbed from oil solution to an oil–air interface, but show significant adsorption at the oil–water interface when dissolved in oil.

From this discussion it is possible also to conclude that adsorption from aqueous solution requires a lower solute concentration to obtain the same level of adsorption if the hydrophobic chain length is increased or if the polar portion of the molecule is less hydrophilic. On the other hand, adsorption from nonpolar solvents is favored when the solute is quite polar.

Since soluble or adsorbed films cannot be compressed, there is no simple direct way to estimate the number of molecules/unit area coming to the surface under a given set of conditions. For relatively simple systems it is possible to estimate this value by application of the Gibbs equation, which relates surface concentration to the surface-tension change produced at different solute activities. The derivation of this equation is beyond the scope of this discussion, but it arises from a classical thermodynamic treatment of the change in free energy when molecules concentrate at the boundary between two phases. The equation may be expressed as

$$\Gamma = -\frac{a}{RT}\frac{d\gamma}{da} \tag{29}$$

where Γ is the moles of solute adsorbed/unit area, R is the gas constant, T is the absolute temperature and $d\gamma$ is the change in surface tension with a change in solute activity, da, at activity a. For dilute solutions of nonelectrolytes, or for electrolytes when the Debye–Hückel equation for activity coefficient is applicable, the value of a may be replaced by solute concentration, c. Since the term dc/c is equal to $d \ln c$, the Gibbs equation is often written as

$$\Gamma = -\frac{1}{RT}\frac{d\gamma}{d \ln c} \tag{30}$$

In this way the slope of a plot of γ vs $\ln c$ multiplied by $1/RT$ should give Γ at a particular value of c. Figure 19-13 depicts typical plots for a series of water-soluble surface-active agents differing only in the alkyl chain length. Note the

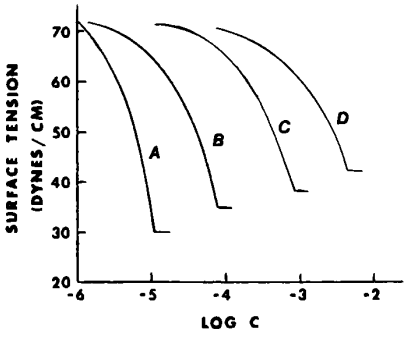


Fig 19-13. The effect of increasing chain length on the surface activity of a surfactant at the air–aqueous solution interface (each figure depicted to differ by two methylene groups with *A*, the longest chain, and *D*, the shortest).

greater reduction of surface tension that occurs at lower concentrations for longer chain-length compounds. In addition, note the greater slopes with increasing concentration, indicating more adsorption (Eq 30), and the abrupt leveling of surface tension at higher concentrations. This latter behavior reflects the self-association of surface-active agent to form micelles which exhibit no further tendency to reduce surface tension. The topic of micelles will be discussed later on page 268.

If one plots the values of surface concentration, Γ, vs concentration, c, for substances adsorbing to the vapor-liquid and liquid-liquid interfaces, using data such as those given in Fig 19-13, one generally obtains an adsorption isotherm shaped like those in Fig 19-9 for vapor adsorption. Indeed, it can be shown that the Langmuir equation (Eq 25) can be fitted to such data when written in the form

$$\Gamma = \frac{\Gamma_{max}\, k'c}{1 + k'c} \tag{31}$$

where Γ_{max} is the maximum surface concentration attained with increasing concentration and k' is related to k in Eq 25. Combining Eqs 29 and 31 leads to a widely used relationship between surface tension change Π (see Eq 28) and solute concentration, c, known as the Syszkowski equation:

$$\Pi = \Gamma_{max} RT \ln (1 + k'c) \tag{32}$$

Mixed Films

It would seem reasonable to expect that the properties of a surface film could be varied greatly if a mixture of surface-active agents were in the film. As an example, consider that a mixture of short- and long-chain fatty acids would be expected to show a degree of "condensation" varying from the "gaseous" state, when the short-chain substance is used in high amount, to a highly condensed state when the longer chain substance predominates. Thus, each component in such a case would operate independently by bringing a proportional amount of film behavior to the system.

More often, the ingredients of a surface film do not behave independently, but, rather, interact to produce a new surface film. An obvious example would be the combination of organic amines and acids which are oppositely charged and would be expected to interact strongly.

In addition to such polar-group interactions, chain–chain interaction will strongly favor mixed condensed films. An important example of such a case occurs when a long-chain alcohol is introduced along with an ionized long-chain substance. Together the molecules form a highly condensed film despite the presence of a high number of like charges. Presumably this occurs as seen in Fig 19-14, by arranging the molecules so that ionic groups alternate with alcohol groups; however, if chain–chain interactions are not strong, the ionic species often will be displaced by the more nonpolar unionized species and "desorb" into the bulk solution.

On the other hand, sometimes the more soluble surface-active agent produces surface pressures in excess of the collapse pressure of the insoluble film and displaces it from the surface. This is an important concept because it is the underlying principle behind cell lysis by surface-active agents and some drugs, and behind the important process of detergency.

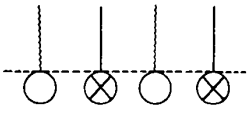


Fig 19-14. A mixed monomolecular film. ⊗: a long-chain ion; O: a long-chain nonionic compound.

Adsorption on Solid Surfaces From Solution

Adsorption to solid surfaces from solution may occur if dissolved molecules and the solid surface have chem groups capable of interacting. Nonspecific adsorption will occur if the solute is surface active and if the surface of the solid is high. This latter case would be the sam occurs at the vapor-liquid and liquid-liquid interfaces. with adsorption to liquid interfaces, adsorption to solid faces from solution generally leads to a monomolecular er, often described by the Langmuir equation or by empirical, yet related, Freundlich equation

$$x/M = kc^n$$

where x is the grams of solute adsorbed by M grams of s in equilibrium with a solute concentration of c. The terr and n are empirical constants. However, as Giles[8] pointed out, the variety of combinations of solutes and ids, and, hence the variety of possible mechanisms of ads tion, can lead to a number of more complex isotherms. particular, adsorption of surfactants and polymers, of g importance in a number of pharmaceutical systems, is not well understood on a fundamental level, and ma some situations even be multilayered.

Adsorption from solution may be measured by separa solid and solution and either estimating the amount of sorbate adhering to the solid or the loss in concentratio adsorbate from solution.

In view of the possibility of solvent adsorption, the la approach really only gives an apparent adsorption. example, if solvent adsorption is great enough, it is poss to end up with an increased concentration of solute a contact with the solid; here, the term negative adsorptic used.

Solvent not only influences adsorption by competing the surface but, as discussed in connection with adsorp at liquid surfaces, the solvent will determine the esca tendency of a solute; eg, the more polar the molecule, the the adsorption that occurs from water. This is seen in 19-15, where adsorption of various fatty acids from w onto charcoal increases with increasing alkyl chain lengt nonpolarity. It is difficult to predict these effects bu general, the more chemically unlike the solute and sol and the more alike the solid surface groups and solute,

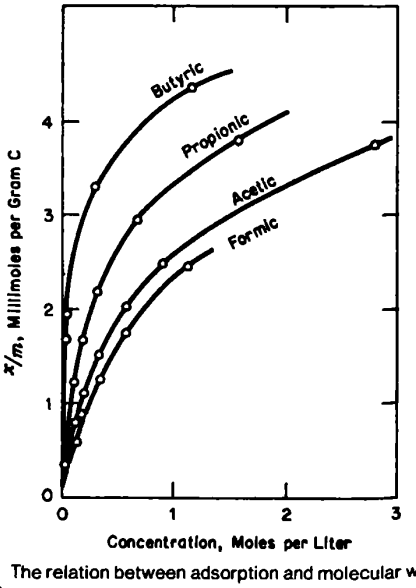


Fig 19-15. The relation between adsorption and molecular weig fatty acids.[9]

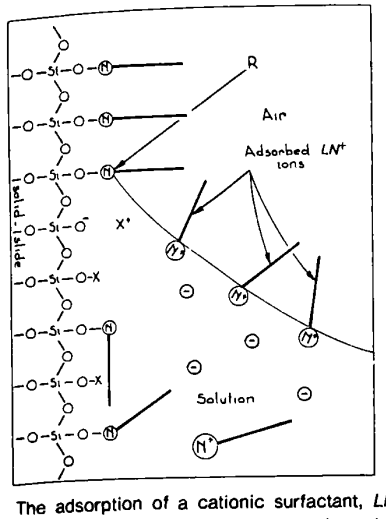


Fig 19-16. The adsorption of a cationic surfactant, LN^+, onto a negatively charged silica or glass surface, exposing a hydrophobic surface as the solid is exposed to air.[10]

greater the extent of adsorption. Another factor which must be kept in mind is that charged solid surfaces, such as polyelectrolytes, will strongly adsorb oppositely charged solutes. This is similar to the strong specific binding seen in gas chemisorption and it is characterized by significant monolayer adsorption at very low concentrations of solute. See Fig 19-16 for an example of such adsorption.

Surface-Active Agents

Throughout the discussion so far, examples of surface-active agents (surfactants) have been restricted primarily to fatty acids and their salts. It has been shown that both a hydrophobic portion (alkyl chain) and a hydrophilic portion (carboxyl and carboxylate groups) are required for their surface activity, the relative degree of polarity determining the tendency to accumulate at interfaces. It now becomes important to look at some of the specific types of surfactants available and to see what structural features are required for different pharmaceutical applications.

The classification of surfactants is quite arbitrary, but one based on chemical structure appears best as a means of introducing the topic. It is generally convenient to categorize surfactants according to their polar portions since the nonpolar portion is usually made up of alkyl or aryl groups. The major polar groups found in most surfactants may be divided as follows: anionic, cationic, amphoteric and nonionic. As we shall see, the last group is the largest and most widely used for pharmaceutical systems, so that it will be emphasized in the discussion that follows.

Types

Anionic Agents—The most commonly used anionic surfactants are those containing carboxylate, sulfonate, and sulfate ions. Those containing carboxylate ions are known as soaps and are generally prepared by the saponification of natural fatty acid glycerides in alkaline solution. The most common cations associated with soaps are sodium, potassium, ammonium, and triethanolamine, while the chain length of the fatty acids ranges from 12 to 18.

The degree of water solubility is greatly influenced by the length of the alkyl chain and the presence of double bonds. For example, sodium stearate is quite insoluble in water at room temperature, whereas sodium oleate under the same conditions is quite water soluble.

Table VII—Effect of Aerosol OT Concentration on the Surface Tension of Water and the Contact Angle of Water with Magnesium Stearate

Concentration, $m \times 10^6$	γ_{sv}	θ
1.0	60.1	120°
3.0	49.8	113°
5.0	45.1	104°
8.0	40.6	89°
10.0	38.6	80°
12.0	37.9	71°
15.0	35.0	63°
20.0	32.4	54°
25.0	29.5	50°

Multivalent ions, such as calcium and magnesium, produce marked water insolubility, even at lower alkyl chain lengths; thus, soaps are not useful in hard water which is high in content of these ions. Soaps, being salts of weak acids, are subject also to hydrolysis and the formation of free acid plus hydroxide ion, particularly when in more concentrated solution.

To offset some of the disadvantages of soaps, a number of long-alkyl-chain sulfonates, as well as alkyl aryl sulfonates such as sodium dodecylbenzene sulfonate, may be used; the sulfonate ion is less subject to hydrolysis and precipitation in the presence of multivalent ions. A popular group of sulfonates, widely used in pharmaceutical systems, are the dialkyl sodium sulfosuccinates, particularly sodium bis-(2-ethylhexyl)sulfosuccinate, best known as Aerosol OT or docusate sodium. This compound is unique in that it is both oil and water soluble and hence forms micelles in both phases. It reduces surface and interfacial tension to low values and acts as an excellent wetting agent in many types of solid dosage forms (see Table VII).

A number of alkyl sulfates are available as surfactants, but by far the most popular member of this group is sodium lauryl sulfate, which is widely used as an emulsifier and solubilizer in pharmaceutical systems. Unlike the sulfonates, sulfates are susceptible to hydrolysis which leads to the formation of the long-chain alcohol, so that pH control is most important for sulfate solutions.

Cationic Agents—A number of long-chain cations, such as amine salts and quaternary ammonium salts, are often used as surface-active agents when dissolved in water; however, their use in pharmaceutical preparations is limited to that of antimicrobial preservation rather than as surfactants. This arises because the cations adsorb so readily at cell membrane structures in a nonspecific manner, leading to cell lysis (eg, hemolysis), as do anionics to a lesser extent. It is in this way that they act to destroy bacteria and fungi.

Since anionic and nonionic agents are not as effective as preservatives, one must conclude that the positive charge of these compounds is important; however, the extent of surface activity has been shown to determine the amount of material needed for a given amount of preservation. Quaternary ammonium salts are preferable to free amine salts since they are not subject to effect by pH in any way; however, the presence of organic anions such as dyes and natural polyelectrolytes is an important source of incompatibility and such a combination should be avoided.

Amphoteric Agents—The major group of molecules falling into this category are those containing carboxylate or phosphate groups as the anion and amino or quaternary ammonium groups as the cation. The former group is represented by various polypeptides, proteins, and the alkyl betaines, while the latter group consist of natural phospholipids such as the lecithins and cephalins. In general, long-chain amphoterics which exist in solution in zwitterionic form are

more surface-active than ionic surfactants having the same hydrophobic group since in effect the oppositely charged ions are neutralized. However, when compared to nonionics, they appear somewhere between ionic and nonionic.

Nonionic Agents—The major class of compounds used in pharmaceutical systems are the nonionic surfactants since their advantages with respect to compatibility, stability, and potential toxicity are quite significant. It is convenient to divide these compounds into those that are relatively water insoluble and those that are quite water soluble.

The major type of compounds making up this first group are the long-chain fatty acids and their water-insoluble derivatives. These include (1) fatty alcohols such as lauryl, cetyl (16 carbons) and stearyl alcohols; (2) glyceryl esters such as the naturally occurring mono-, di- and triglycerides; and (3) fatty acid esters of fatty alcohols and other alcohols such as propylene glycol, polyethylene glycol, sorbitan, sucrose and cholesterol. Included also in this general class of nonionic water-insoluble compounds are the free steroidal alcohols such as cholesterol.

To increase the water solubility of these compounds and to form the second group of nonionic agents, polyoxyethylene groups are added through an ether linkage with one of their alcohol groups. The list of derivatives available is much too long to cover completely, but a few general categories will be given.

The most widely used compounds are the polyoxyethylene sorbitan fatty acid esters which are found in both internal and external pharmaceutical formulations. Closely related compounds include polyoxyethylene glyceryl, and steroidal esters, as well as the comparable polyoxypropylene esters. It is also possible to have a direct ether linkage with the hydrophobic group as with a polyoxyethylene–stearyl ether or a polyoxyethylene–alkyl phenol. These ethers offer advantages since, unlike the esters, they are quite resistant to acidic or alkaline hydrolysis.

Besides the classification of surfactants according to their polar portion, it is useful to have a method that categorizes them in a manner that reflects their interfacial activity and their ability to function as wetting agents, emulsifiers, solubilizers, etc. Since variation in the relative polarity or nonpolarity of a surfactant significantly influences its interfacial behavior, some measure of polarity or nonpolarity should be useful as a means of classification.

One such approach assigns a hydrophile–lipophile balance number (HLB) for each surfactant and, although developed by a commercial supplier of one group of surfactants, the method has received wide-spread application. The HLB value, as originally conceived for nonionic surfactants, is merely the percentage weight of the hydrophilic group divided by five in order to reduce the range of values. On a molar basis, therefore, a 100% hydrophilic molecule (polyethylene glycol) would have a value of 20.

Thus, an increase in polyoxyethylene chain length increases polarity and, hence, the HLB value; at constant polar chain length, an increase in alkyl chain length or number of fatty acid groups decreases polarity and the HLB value. One immediate advantage of this system is that to a first approximation one can compare any chemical type of surfactant to another type when both polar and nonpolar groups are different.

HLB values for nonionics are calculable on the basis of the proportion of polyoxyethylene chain present; however, in order to determine values for other types of surfactants it is necessary to compare physical chemical properties reflecting polarity with those surfactants having known HLB values.

Relationships between HLB and phenomena such as water solubility, interfacial tension, and dielectric constant have been used in this regard. Those surfactants exhibiting values greater than 20 (eg, sodium lauryl sulfate) demonstrate hydrophilic behavior in excess of the polyoxyethylene groups alone. Table XIX, page 304, presents HLB values for a variety of surface-active agents.

Surfactant Properties in Solution and Micelle Formation

As seen in Fig 19-13, increasing the concentration of surface-active agents in aqueous solution causes a decrease in the surface tension of the solution until a certain concentration where it then becomes essentially constant with increasing concentration. That this change is associated with changes also taking place in the bulk solution rather than just at the surface can be seen in Fig 19-17, which shows the same abrupt change in bulk solution properties such as solubility, equivalent conductance and osmotic pressure as with surface properties. The most reasonable explanation for these effects is that the solute molecules self-associate to form soluble aggregates which exhibit markedly different properties from the monomers in solution. Such aggregates (Fig 19-18*A*) appear to exhibit no tendency to adsorb to the surface since the surface and interfacial tension above this solute concentration do not change to any significant extent. Such aggregates, known as micelles, form over such a very narrow range of concentrations that one can speak of a critical micellization concentration (cmc). These micelles form for essentially the same reasons that cause molecules to be adsorbed; the lack of affinity of the hydrophobic chains for water molecules and the tendency for strong hydrophobic chain-chain interactions when the chains are oriented closely together in the micelle, coupled with the gain in entropy due to the loss of the ice-like structure of water when the chains are separated from water, lead to a favorable free energy change for micellization. The longer the hydrophobic chain or the less the polarity of the polar group, the greater the tendency for monomers to "escape" from the water to form micelles and, hence the lower the cmc (see Fig 19-13).

In dilute solution (still above the cmc) the micelles can be considered to be approximately spherical in shape (Fig 19-18*A* and *B*), while at higher concentrations they become

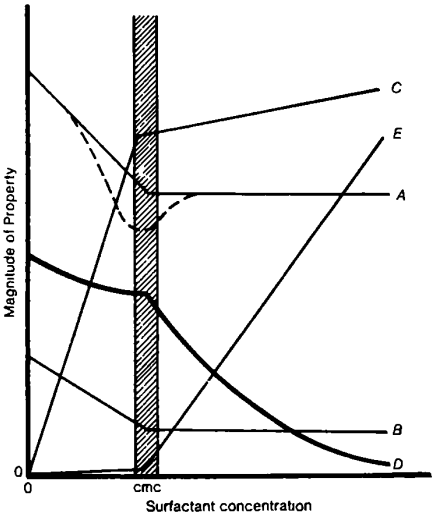


Fig 19-17. Effect of surfactant concentration and micelle formation on various properties of the aqueous solution of an ionic surfactant. *A:* Surface tension; *B:* interfacial tension; *C:* osmotic pressure; *D:* equivalent conductivity; *E:* solubility of compound with very low solubility in pure water [11]

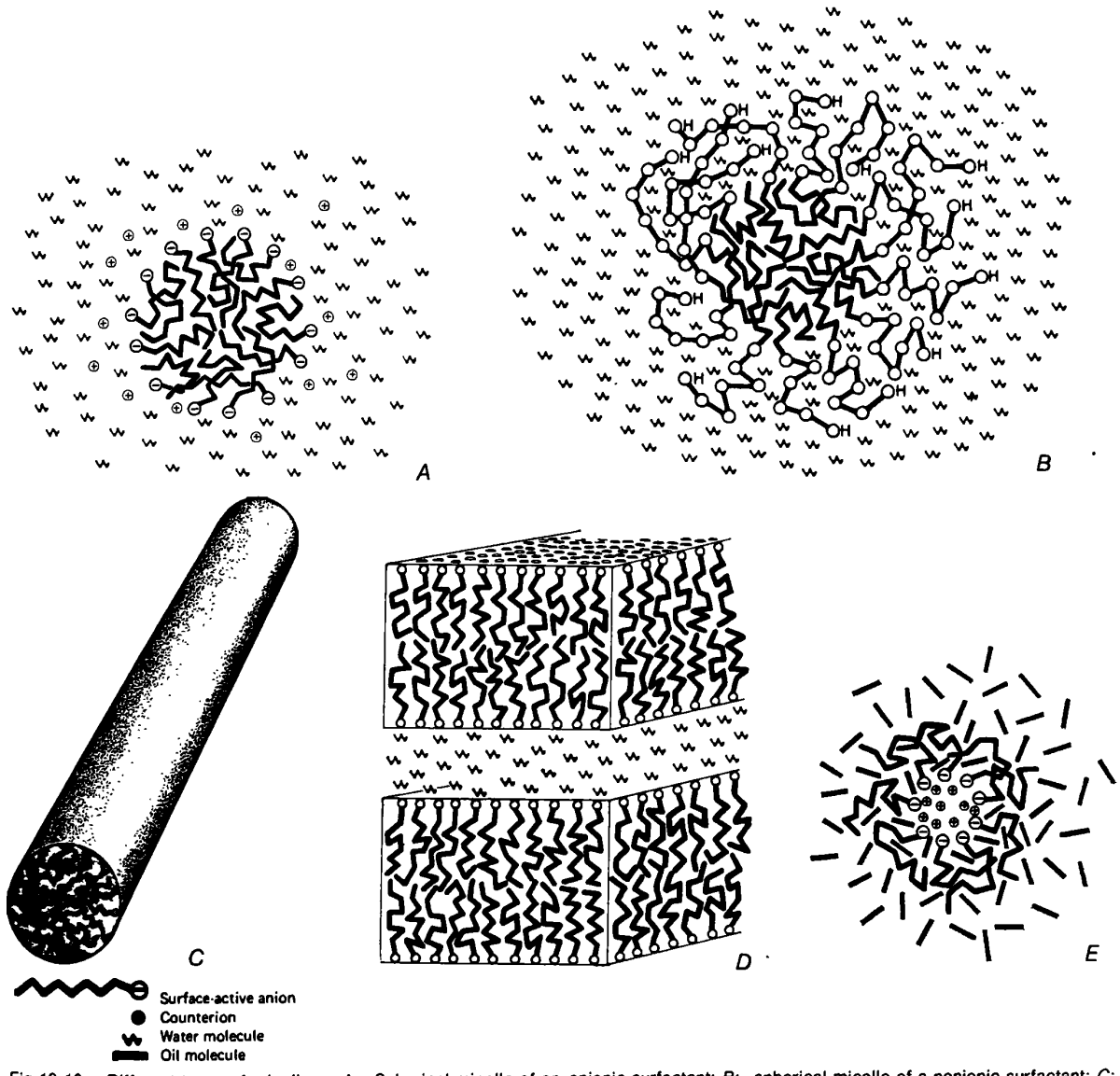


Fig 19-18. Different types of micelles. *A:* Spherical micelle of an anionic surfactant; *B:* spherical micelle of a nonionic surfactant; *C:* cylindrical micelle of an ionic surfactant; *D:* lamellar micelle of an ionic surfactant; *E:* reverse micelle of an anionic surfactant in oil.[11]

more asymmetric and eventually assume cylindrical (Fig 19-18*C*) or lamellar (Fig 19-18*D*) structures. It is important to recognize that equilibrium, and hence reversibility, exists between the monomers and the various types of micelles. The sizes of such micelles depend on the number of monomers per micelle and the size and molecular shape of the individual monomers. In Table VIII are given the cmc and number of monomers per micelle for different types of surfactants. Note for the nonionic surfactants that the longer the polyoxyethylene chain, and hence the more polar and bulkier the molecule, the higher the cmc, ie the less the tendency for micelle formation. It is also possible for oil-soluble surfactants to show a tendency to self-associate into "reverse micelles in nonpolar solvents, as depicted in Fig 19-18*E*, with their polar groups all oriented away from the solvent. In general these micelles tended to be smaller and to aggregate over a wider range of concentrations than seen in water, and therefore, to exhibit no well-defined cmc.

Micellar Solubilization

As seen in Fig 19-18, the interior of surfactant micelles formed in aqueous media consists of hydrocarbon "tails" in liquid-like disorder. The micelles, therefore, resemble miniscule pools of liquid hydrocarbon surrounded by shells of polar "head groups." Compounds which are poorly soluble in water but soluble in hydrocarbon solvents, can be dissolved inside these micelles, ie, they are brought homogeneously into an overall aqueous medium.

Being hydrophobic and oleophilic, the solubilized molecules are located primarily in the hydrocarbon core of the micelles (see Fig 19-19*A*). Even water-insoluble drugs usually contain polar functional groups such as hydroxyl, carbonyl, ether, amino, amide, and cyano. Upon solubilization, these hydrophilic groups locate on the periphery of the micelle among the polar headgroups of the surfactant in order to become hydrated (see Fig 19-19*B*). For instance,

Table VIII—Critical Micelle Concentrations and Micellar Aggregation Numbers of Various Surfactants in Water at Room Temperature

Structure	Name	CMC, m*M*/L	Surfactant molecules/micelle
n-$C_{11}H_{23}COOK$	Potassium laurate	24	50
n-$C_8H_{17}SO_3Na$	Sodium octant sulfonate	150	28
n-$C_{10}H_{21}SO_3Na$	Sodium decane sulfonate	40	40
n-$C_{12}H_{25}SO_3Na$	Sodium dodecane sulfonate	9	54
n-$C_{12}H_{25}OSO_3Na$	Sodium lauryl sulfate	8	62
n-$C_{12}H_{25}OSO_3Na$	Sodium lauryl sulfate[a]	1	96
	Sodium di-2-ethylhexyl sulfosuccinate	5	48
n-$C_{10}H_{21}N(CH_3)_3Br$	Decyltrimethylammonium bromide	63	36
n-$C_{12}H_{25}N(CH_3)_3Br$	Dodecyltrimethylammonium bromide	14	50
n-$C_{14}H_{29}N(CH_3)_3Br$	Tetradecyltrimethylammonium bromide	3	75
n-$C_{14}H_{29}N(CH_3)_3Cl$	Tetradecyltrimethylammonium chloride	3	64
n-$C_{12}H_{25}NH_3Cl$	Dodecylammonium chloride	13	55
n-$C_{12}H_{25}O(CH_2CH_2O)_8H$	Octaoxyethylene glycol monododecyl ether	0.13	132
n-$C_{12}H_{25}O(CH_2CH_2O)_8H$[b]		0.10	301
n-$C_{12}H_{25}(CH_2CH_2O)_{12}H$	Dodecaoxyethylene glycol monododecyl ether	0.14	78
n-$C_{12}H_{25}O(CH_2CH_2O)_{12}H$[b]		0.091	116
t-C_8H_{17}-C_6H_4-$O(CH_2CH_2O)_{9.7}H$	Decaoxyethylene glycol mono-p,t-octylphenyl ether (octoxynol 9)	0.27	100

[a] Interpolated for physiologic saline, 0.154 *M* NaCl.
[b] At 55° instead of 20°.

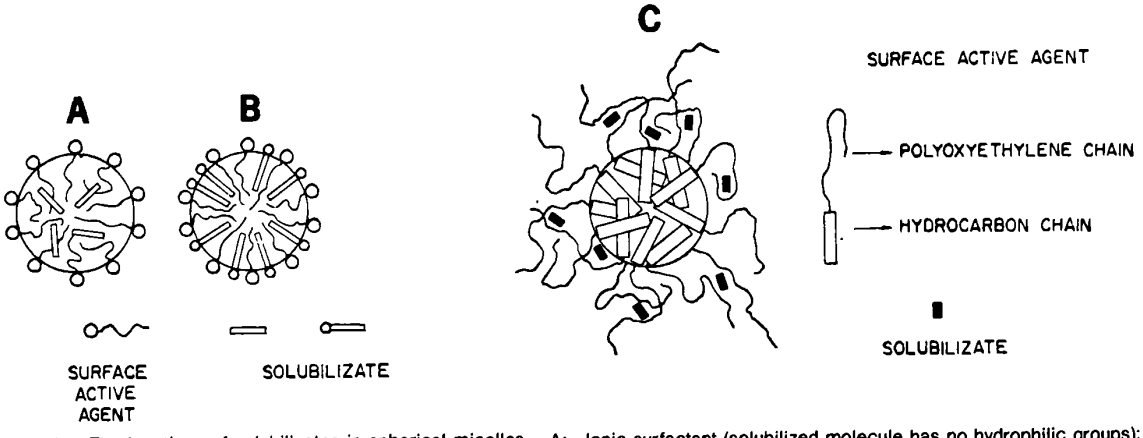


Fig 19-19. The locations of solubilizates in spherical micelles. A: Ionic surfactant (solubilized molecule has no hydrophilic groups); B: ionic surfactant (solubilized molecule has a hydrophilic group); C: nonionic surfactant (polar solubilizate) [12]

when cholesterol or dodecanol is solubilized by sodium lauryl sulfate, their hydroxyl groups penetrate between sulfate ions and are even bound to them by hydrogen bonds, while their hydrocarbon portions are immersed among the dodecyl tails of the surfactant which make up the core of the micelle.

Micelles of polyoxyethylated nonionic surfactants consist of an outer shell of hydrated polyethylene glycol moieties and a core of hydrocarbon moieties. Compounds like phenol, cresol, benzoic acid, salicylic acid, and esters of *p*-hydroxy and *p*-aminobenzoic acids have some solubility in water and in oils but considerable solubility in liquids of intermediate polarity like ethanol, propylene glycol or aqueous solutions of polyethylene glycols. When solubilized by nonionic micelles, they are located in the hydrated outer polyethylene glycol shell as shown in Fig 19-19*C*. Since these compounds have hydroxyl or amino groups, they frequently form complexes with the ether oxygens of the surfactant by hydrogen bonding.

Solubilization is generally nonspecific: any drug which is appreciably soluble in oils can be solubilized. Each has a solubilization limit, comparable to a limit of solubility, which depends on temperature and on the nature and concentration of the surfactant. Hartley distinguishes two categories of solubilizates. The first consists of comparatively large, asymmetrical and rigid molecules forming crystalline solids, such as steroids and dyes. These do not blend in with the normal paraffin tails which make up the micellar core; because of dissimilarity in structure, they remain distinct as solute molecules. They are sparingly solubilized by surfactant solutions, a few molecules/micelle at saturation (see Table IX). The number of carbon atoms in the micellar hydrocarbon core required to solubilize a molecule of steroid or dye at saturation is of the same order of magnitude as the number of carbon atoms of bulk liquid dodecane or hexadecane per molecule of steroid or dye in their saturated solutions in these liquids.

Since solubilization depends on the presence of micelles, it does not take place below the cmc. It can, therefore, be used to determine the cmc, particularly when the solubilizate is a dye or another compound easy to assay. Plotting the maximum amount of a water-insoluble dye solubilized by aqueous surfactant, or the absorbance of its saturated solutions, versus the surfactant concentration produces a straight line which intersects the surfactant concentration axis at the cmc. Above the cmc, the amount of solubilized dye is directly proportional to the number of micelles and, therefore,

Table IX—Micellar Solubilization Capacities of Different Surfactants for Estrone[13]

Surfactant	Concentration range, molarity	Temp, °C	Moles surfactant/ mole solubilized estrone
Sodium laurate	0.025–0.023	40	91
Sodium oleate	0.002–0.35	40	53
Sodium lauryl sulfate	0.004–0.15	40	71
Sodium cholate	0.09–0.23	20	238
Sodium deoxycholate	0.007–0.36	20	476
Diamyl sodium sulfosuccinate	0.08–0.4	40	833
Dioctyl sodium sulfosuccinate	0.002–0.05	40	196
Tetradecyltrimethylammonium bromide	0.005–0.08	20	45
Hexadecylpyridinium chloride	0.001–0.1	20	32
Polysorbate 20	0.002–0.15	20	161
Polysorbate 60	0.0008–0.11	20	83

proportional to the overall surfactant concentration. Below the cmc, no solubilization takes place. This is represented by Curve E of Fig 19-17.

The second category of compounds to be solubilized are often liquid at room temperature and consist of relatively small, symmetrical, and/or flexible molecules such as many constituents of essential oils. These molecules mix and blend in freely with the hydrocarbon portions of the surfactants in the core of the micelles, so as to become indistinguishable from them. Such compounds are extensively solubilized and in the process usually swell the micelles: they augment the volume of the hydrocarbon core and increase the number of surfactant molecules per micelle. Their solubilization frequently lowers the cmc.

Microemulsions[14-16]

Microemulsions are liquid dispersions of water and oil that are made homogeneous, transparent, and stable by the addition of relatively large amounts of a surfactant and a cosurfactant. *Oil* is defined as a liquid of low polarity and low miscibility with water, eg, toluene, cyclohexane, mineral or vegetable oils.

Microemulsions are intermediate in properties between micelles containing solubilized oils and emulsions. While emulsions are lyophobic and unstable, microemulsions are on the borderline between lyophobic and lyophilic colloids. True microemulsions are thermodynamically stable.[17] Therefore, they are formed spontaneously when oil, water, surfactants, and cosurfactants are mixed together. The unstable emulsions require input of considerable mechanical energy for their preparation, which may be supplied by colloid mills, homogenizers or ultrasonic generators.

Both emulsions and microemulsions may contain high volume fractions of the internal phase. For instance, some O/W systems contain 75% (v/v) of oil dispersed in 25% water, although lower internal phase volume fractions are more common.

At low surfactant concentrations, viz, low multiples of the cmc, micelles are spheres (Fig 19-18*A*, *B* and *E*) or ellipsoids. When an oil is solubilized by micelles in water, it blends into the micellar core formed by the hydrocarbon tails of the surfactant molecules (Fig 19-19) and swells the micelles.

Spherical or ellipsoidal micelles are nearly monodisperse, and their mean diameters are in the range of 25 to 60 Å. Microemulsion droplets also have a narrow droplet size distribution with a mean diameter range of approximately 60 to 1000 Å. Since the droplet diameters are less than ¼ of the wavelength of light (4200 Å for violet and 6600 Å for red light), microemulsions scatter little light and are, therefore, transparent or at least translucent.

Emulsions have very broad droplet size distributions. Only the smallest droplets, with diameters of about 1000 to 2000 Å, are below the resolving power of the light microscope. The upper size limit is 25 or 50 μm (250,000 or 500,000 Å). Because emulsion droplets are comparable in size, or larger than the wavelength of visible light, they scatter it more or less strongly depending on the difference in refractive index between oil and water. Thus, most emulsions are opaque.

The three disperse systems—micellar solutions, microemulsions, and emulsions—can be of the O/W (oil-in-water) or W/O type. Aqueous micellar surfactant solutions can solubilize oils and lipid-soluble drugs in the core formed by their hydrocarbon chains. Likewise, oil-soluble surfactants like sorbitan monooleate and docusate sodium form "reverse micelles" in oils (Fig 19-18*E*) capable of solubilizing water in the polar center. The solubilized oil in the former micelles and the solubilized water in the latter may in turn enhance the micellar solubilization of oil-soluble and water-soluble drugs, respectively.

Oil-soluble drugs have been incorporated into O/W emulsions by dissolving them in the oil phase before emulsification.[18] By the same token, it may be possible to dissolve oil-soluble drugs in a vegetable oil and make an oral or parenteral O/W microemulsion. The advantage of such microemulsion systems over conventional emulsions is their smaller droplet size and superior shelf stability. Aqueous micellar solutions[19] and O/W microemulsions[20] have both been used as aqueous reaction media for oil-soluble compounds.

Emulsions and micellar solutions of oils solubilized in aqueous surfactant solutions consist of three components, oil, water and surfactant. Microemulsions generally require a fourth component, called *cosurfactant*. Commonly used cosurfactants are linear alcohols of medium chain length, which are sparingly miscible with water. Since the cosurfactants as well as the surfactants are surface-active, they promote the generation of extensive interfaces through the spontaneous dispersion of oil in water, or vice-versa, resulting in the formation of microemulsions. The large interfacial area between oil and water permits the extensive formation of a mixed interfacial film consisting of surfactant and cosurfactant. This film is called the "interphase" because it is thicker than the surfactant monolayers formed at oil-water interfaces in emulsions. The interfacial tension at the oil-water interface in microemulsions approaches zero, which also contributes to their spontaneous formation. According to another viewpoint, microemulsions are regarded as micelles extensively swollen by large amounts of solubilized oil.

Typical formulations for an O/W and a W/O microemulsion are shown in Table X. The ratio, g surfactant/g solubilized or emulsified oil or water is in the range of 2 to 20 for micellar solutions and 0.01 to 0.1 for emulsions. Microemulsions have intermediate values: The ratios for the formulations in Table X are near unity. In industrial formulations,

Table X—Microemulsion Formulations

Compound	Function	Content in microemulsions, % O/W	Content in microemulsions, % W/O
Sodium lauryl sulfate	Surfactant	13	10
1-Pentanol	Cosurfactant	8	25
Xylene	Oil	8	50
Water		71	15

the ratios are closer to 0.1 to reduce costs. Microemulsions are used in such diverse applications as floor polish and agricultural pesticide formulations and in tertiary petroleum recovery. The use of O/W microemulsions as aqueous vehicles for oil-soluble drugs to be administered by the percutaneous, oral or parenteral route is being investigated.

Colloidal Dispersions

Historical Background of Colloids

The term *colloid*, derived from the Greek word for glue, was applied *ca* 1850 by the British chemist Thomas Graham to polypeptides such as albumin and gelatin, to vegetable gums such as acacia, starch and dextrin, and to inorganic compounds such as gelatinous metal hydroxides and Prussian blue (ferric ferrocyanide). These compounds did not crystallize, and diffused very slowly when dissolved or dispersed in water. They could be separated from ordinary solutes such as salts and sugar, called "crystalloids," as the latter diffused through the fine pores of dialysis membranes made from animal gut which retained the "colloids." "Crystalloids" crystallized readily from solution.[21,22]

Von Weimarn was the first to identify colloidality as a state of subdivision of matter rather than as a category of substances. Many of Graham's "colloids," especially proteins, have been crystallized. Moreover, von Weimarn was able to prepare all "crystalloids" investigated in the colloidal state. Colloidal dispersions by the condensation method resulted from high relative supersaturation, which produced a large number of small nuclei.[21-23,28] For instance, clear, transparent solidified jellies were prepared by cooling aqueous solutions of $CaCl_2$, $Ba(SCN)_2$ and $Al_2(SO_4)_3$, and aqueous-alcoholic solutions of NaCl, KCl, NH_4Cl, KSCN, NaBr and NH_4NO_3 which were nearly saturated at room temperature.[28]

Colloid chemistry became a science in its own right around 1906, when Wolfgang Ostwald wrote the booklet "The World of the Neglected Dimensions." In it, he focused on colloidal systems as a state of matter that has disperse phases intermediate in size between small molecules or ions in solution and large, visible particles in suspension. Ostwald became the first editor of the journal *Kolloid-Zeitschrift* in 1907. The studies of colloidal systems and surface or interfacial phenomena are intimately related. The properties of colloidal dispersions are largely governed by the nature of the surface of their particles. The division of the American Chemical Society specializing in colloidal systems and interfaces is called the "Division of Colloid and Surface Chemistry," while the pertinent session of the Gordon Research Conferences is called "Chemistry at Interfaces."

Colloid and surface chemistry deals with an unusually wide variety of industrial and biological systems. A few examples are catalysts, lubricants, adhesives, latexes for paints, rubbers and plastics, soaps and detergents, clays, packaging films, cigarette smoke, liquid crystals, cell membranes, mucous secretions and aqueous humors.

Definitions and Classifications

Colloidal Systems and Interfaces

Colloidal dispersions consist of at least two discrete phases, namely, one or more disperse, dispersed or internal phases and a continuous or external phase called the *dispersion medium* or *vehicle*. What distinguishes colloidal dispersions from solutions and coarse dispersions is the particle size of the disperse phase. Systems in the colloidal state contain one or more substances that have at least one dimension in the range of 10 to 100 Å (1 Angstrom unit = 10^{-8} cm = 10^{-10} m) or 1–10 nm (1 nanometer = 10^{-9} m) at the lower end, and a few micrometers (μm) at the upper end (1 μm = 10^4 Å = 10^{-6} m). Thus blood, cell membranes, the thinner nerve fibers, milk, rubber latex, fog and beer foam are colloidal systems. Some types of materials, such as many emulsions, and oral suspensions of most organic drugs, are coarser than true colloidal systems but exhibit similar behavior. Even though serum albumin, acacia and povidone form true or molecular solutions in water, the size of the individual solute molecules places such solutions in the colloidal range (particle size > 10 Å).[21-27]

The following features distinguish colloidal dispersions from coarse suspensions. Disperse particles in the colloidal range are usually too fine to be visible in a light microscope, because at least one dimension measures 1 μm or less. They are often visible in the ultramicroscope and always in the electron microscope. Coarse suspended particles are frequently visible to the naked eye and always in the light microscope. Colloidal particles, as opposed to coarse particles, pass through ordinary filter paper but are retained by dialysis or ultrafiltration membranes. Because of their small size, colloidal dispersions undergo little or no sedimentation or creaming: Brownian motion maintains the disperse particles in suspension (see below).

Except for high polymers, most soluble substances can be prepared either as low-molecular-weight solutions, or as colloidal dispersions or coarse suspensions depending on the choice of the dispersion medium and the dispersion technique.[26,28]

Because of the small size of colloidal particles, appreciable fractions of their atoms, ions or molecules are located in the boundary layer between a particle and air (surface) or between a particle and a liquid or solid (interface). The ions in the surface of a sodium chloride crystal and the water molecules in the surface of a rain drop are subjected to unbalanced forces of attraction, whereas the ions or molecules in the interior of the materials are surrounded by similar ions or molecules on all sides, with balanced force fields. Thus a surface free energy component is added to the total free energy of colloidal particles, which becomes relatively more important as the particles become smaller, ie, as greater fractions of their ions, atoms or molecules are located in their surface or interfacial region. Hence the solubility of very fine solid particles and the vapor pressure of very small liquid droplets are larger than the corresponding values of coarse particles and large drops of the same materials, respectively.

Specific Surface Area—Decreasing particle size increases the surface-to-volume ratio, which is expressed as the specific surface area A_{sp}, namely, the area A (cm^2) per unit volume V ($1\ cm^3$) or per unit mass M (1 gram). For a sphere, $A = 4\pi r^2$ and $V = 4/3\ \pi r^3$. If the density, d, of the material is expressed in g/cm^3, the specific surface area is

$$A_{sp} = \frac{A}{V} = \frac{4\pi r^2}{4/3\pi r^3} = \frac{3}{r}\ cm^2/cm^3 = \frac{3}{r}\ cm^{-1}$$

or

$$A_{sp} = \frac{A}{M} = \frac{A}{Vd} = \frac{4\pi r^2}{4/3\pi r^3 d} = \frac{3}{rd}\ cm^2/g$$

Table XI—Effect of Comminution on Specific Surface Area of a Volume of $4\pi/3$ cm^3, Divided into Uniform Spheres of Radius R

Number of spheres	R	A_{sp} cm^2/cm^3
1	1 cm	3
10^3	0.1 cm = 1 mm	3×10
10^6	0.1 mm	3×10^2
10^9	0.01 mm = 10 μm	3×10^3
10^{12}	1 μm	3×10^4
10^{15}	0.1 μm	3×10^5
10^{18}	0.01 μm	3×10^6
10^{21}	10 Å = 1 nm	3×10^7
10^{23}	1 Å	3×10^8

Shaded region corresponds to colloidal particle-size range

Table XI illustrates the effect of comminution on the specific surface area of 4 $\pi/3$ cm^3 of a material consisting initially of one sphere of 1 cm radius. As the material is broken up into an increasingly larger number of smaller and smaller spheres, its specific surface area increases commensurately.

The solid adsorbents activated charcoal and kaolin have specific surface areas of about 6×10^6 cm^2/g and 10^4 cm^2/g, respectively. One gram of activated charcoal, because of its extensive porosity and internal voids, has an area equal to ⅙ acre.

In conclusion, colloidal systems by definition are those polyphasic systems where at least one dimension of the disperse phase measures between 10 or 100 Å and a few micrometers. The term "colloidal" designates a state of matter characterized by submicroscopic dimensions rather than certain substances. Any dispersed substance with the proper dimension or dimensions is in the colloidal state.

Physical States of Disperse and Continuous Phases

A useful classification of colloidal systems (systems in the colloidal particle size range) is based on the state of matter of the disperse phase and the dispersion medium, ie, whether they are solid, liquid or gaseous.[25,27] Table XII summarizes the various combinations and lists examples. A *sol* is the colloidal dispersion of a solid in a liquid or gaseous medium. Prefixes designate the dispersion medium, such as hydrosol, alcosol, aerosol for water, alcohol and air, respectively. Sols are fluid. If the solid particles form bridged structures possessing some mechanical strength, the system is called a gel (hydrogel, alcogel, aerogel).

Interaction Between Disperse Phase and Dispersion Medium

A second useful classification of colloidal dispersions, originated by Ostwald, is based on the affinity or interaction between the disperse phase and the dispersion medium.[2,3,8] It refers mostly to solid-in-liquid dispersions. According to this classification, colloidal dispersions are divided into the two broad categories of lyophilic and lyophobic. Some soluble, low-molecular-weight substances have molecules with both tendencies, forming a third category called association colloids.

Lyophilic Dispersions—Where there is considerable attraction between the disperse phase and the liquid vehicle, ie, extensive solvation, the system is said to be *lyophilic* (solvent-loving). If the dispersion medium is water, the system is said to be *hydrophilic*. Such solids as bentonite, starch, gelatin, acacia and povidone swell, disperse or dissolve spontaneously in water.

Hydrophilic colloidal dispersions can be subdivided further as follows:

True solutions, formed by water-soluble polymers (acacia and povidone).

Gelled solutions, gels or jellies if the polymers are present at high concentrations and/or at temperatures where their water solubility is low. Examples of such hydrogels are relatively concentrated solutions of gelatin and starch, which set to gels on cooling, or of methylcellulose, which gel on heating.

Particulate dispersions, where the solids do not form molecular solutions but remain as discrete though minute particles. Bentonite and microcrystalline cellulose form such hydrosols.

Lipophilic or oleophilic substances have pronounced affinity for oils. Oils are nonpolar liquids consisting mainly of hydrocarbons, with few polar groups and low dielectric constants. Examples are mineral oil, benzene, carbon tetrachloride, vegetable oils (cottonseed or peanut oil) and essential oils (lemon or peppermint oil). Substances which form *oleophilic* colloidal dispersions include polymers like polystyrene and unvulcanized or gum rubber, which dissolve molecularly in benzene, magnesium or aluminum stearate or which dissolve or disperse in cottonseed oil, and activated charcoal, which forms sols or particulate dispersions in all oils.

Because of the high affinity or attraction between the dispersion medium and the disperse phase, lyophilic dispersions form spontaneously when the liquid vehicle is brought into contact with the solid phase. They are thermodynamically stable and reversible, ie, they are easily reconstituted even after the dispersion medium has been removed from the solid phase.[22,24-27]

Table XII—Classification of Colloidal Dispersions According to State of Matter

Disperse Phase	Dispersion Medium (Vehicle)		
	Solid	Liquid	Gas
Solid	Zinc oxide paste (zinc oxide + starch in petrolatum). Toothpaste (dicalcium phosphate or calcium carbonate with sodium carboxymethylcellulose binder). Pigmented plastics (titanium dioxide in polyethylene).	Sols: Bentonite Magma NF. Trisulfapyrimidines Oral Suspension USP. Magnesia and Alumina Oral Suspension USP. Tetracycline Oral Suspension USP.	Solid aerosols: Smoke, dust. Epinephrine Bitartrate Inhalation Aerosol USP. Isoproterenol Sulfate Inhalation Aerosol.
Liquid	Absorption bases (aqueous medium in Hydrophilic Petrolatum USP). Emulsion bases (oil in Hydrophilic Ointment USP). Butter.	Emulsions: Mineral Oil Emulsion USP. Soybean oil in water emulsion for IV feeding. Milk. Mayonnaise.	Liquid aerosols: Mist, fog. Nasal relief sprays (naphazoline hydrochloride solution). Betamethasone Valerate Topical Aerosol USP. Povidone-Iodine Topical Aerosol.
Gas	Solid foams (foamed plastics and rubbers). Pumice.	Foams. Carbonated beverages. Effervescent salts in water.	No colloidal dispersions.

Lyophobic Dispersions—When there is little attraction between the disperse phase and the dispersion medium, the dispersion is said to be *lyophobic* (solvent-hating). *Hydrophobic* dispersions consist of particles that are not hydrated, so that water molecules interact with or attract one another in preference to solvating the particles. They include aqueous dispersions of oleophilic materials such as polystyrene or gum rubber (latex), steroids and other organic lipophilic drugs, paraffin wax, magnesium stearate, and of cottonseed or soybean oil (emulsion). While lipophilic materials are generally hydrophobic, materials like sulfur, silver chloride and gold form hydrophobic dispersions without being lipophilic. Water-in-oil emulsions are lyophobic dispersions in lipophilic vehicles.

Because of the lack of attraction between the disperse and the continuous phase, lyophobic dispersions are intrinsically unstable and irreversible. Their large surface free energy is not lowered by solvation. The dispersion process does not take place spontaneously, and once the dispersion medium has been separated from the disperse phase, the dispersion is not easily reconstituted. The dividing line between hydrophilic and hydrophobic dispersions is not very sharp. For instance, gelatinous hydroxides of polyvalent metals such as $Al(OH)_3$ and $Mg(OH)_2$, and clays such as bentonite and kaolin, possess some characteristics of both.[22,24,27]

Association Colloids—Organic compounds which contain large hydrophobic moieties together with strongly hydrophilic groups in the same molecule are said to be amphiphilic. While the individual molecules are generally too small to bring their solutions into the colloidal size range, they tend to associate in aqueous or oil solutions into micelles (see above). Because micelles are large enough to qualify as colloidal particles, such compounds are called association colloids.

Lyophobic Dispersions

Most of the discussion of lyophobic dispersions deals with hydrophobic dispersions or hydrosols (hydrophobic solids or liquids dispersed in aqueous media) because water is the most widely used vehicle. They comprise aqueous dispersions of insoluble organic and inorganic compounds which usually have low degrees of hydration. Organic compounds which are preponderantly hydrocarbon in nature and possess few hydrophilic or polar groups are insoluble in water and hydrophobic.

Hydrophobic dispersions are intrinsically unstable. The most stable state of such systems contains the disperse phase coalesced into large crystals or drops, so that the specific surface area and surface free energy are reduced to a minimum. Therefore, mechanical, chemical or electrical energy must be supplied to the system to break up the disperse phase into small particles, providing for the increase in surface free energy resulting from the parallel increase in specific surface area. Furthermore, special means must be found to stabilize hydrophobic dispersions, preventing the otherwise spontaneous coalescence or coagulation of the disperse phase after it has been finely dispersed.

Preparation and Purification of Lyophobic Dispersions

Colloidal dispersions are intermediate in size between true solutions and coarse suspensions. They can be prepared by aggregation of small molecules or ions until particles of colloidal dimensions result (condensation methods), or by reducing coarse particles to colloidal dimensions through comminution or peptization (dispersion methods).

Dispersion Methods—The first method, *mechanical disintegration* of solids and liquids into small particles and their dispersion in a fluid vehicle, is frequently carried out by input of mechanical energy via shear or attrition. Equipment such as colloid and ball mills, micronizers and, for emulsions, homogenizers is described in Chapters 83 and 88 and in Ref 29. Dry grinding with inert, water-soluble diluting agents also produces colloidal dispersions. Sulfur hydrosols may be prepared by triturating the powder with urea or lactose followed by shaking with water.

Ultrasonic generators provide exceptionally high concentrations of energy. Successful dispersion of solids by means of ultrasonic waves can only be achieved with comparatively soft materials such as many organic compounds, sulfur, talcum, and graphite. Where fine emulsions are mandatory, such as soybean oil-in-water emulsions used for intravenous feeding, emulsification by ultrasound waves is the method of choice.[29] The formation of aerosols is described in Chapter 92.

It should be reiterated that hydrosols of hydrophobic substances are intrinsically unstable. While mechanical disintegration may break up the disperse phase into colloidal particles, the resultant dispersions tend towards separation of that phase. Recrystallization, coagulation or coalescence causes the disperse particles to become progressively coarser and fewer, ultimately resulting in the separation of a macroscopic phase. To avoid this, stabilizing agents must be added during or shortly after the dispersion process (see below). For instance, lecithin may be used to stabilize soybean oil emulsions.

Peptization is a second method for preparing colloidal dispersions. The term, coined by Graham, is defined as the breaking up of aggregates or secondary particles into smaller aggregates or into primary particles in the colloidal size range. Particles which are not formed of smaller ones are called "primary." Peptization is synonymous with *deflocculation*. It can be brought about by the removal of flocculating agents, usually electrolytes, or by the addition of deflocculating or peptizing agents, usually surfactants, water-soluble polymers or ions which are adsorbed at the particle surface.[24,27]

The mechanisms of the following examples are explained in subsequent sections. When powdered activated charcoal is added to water with stirring, the aggregated grains are broken up only incompletely and the resultant suspension is gray and translucent. The addition of 0.1% or less of sodium lauryl sulfate or octoxynol disintegrates the grains into finely dispersed particles forming a deep black and opaque dispersion. Ferric or aluminum hydroxide freshly precipitated with ammonia can be peptized with small amounts of acids which reduce the pH below the isoelectric points of the hydroxides (see below). Even washing the gelatinous precipitate of $Al(OH)_3$ with water tends to peptize it. In quantitative analysis, the precipitate is therefore washed with dilute solutions of ammonium salts that act as flocculating agents, rather than with water.

Condensation Methods—The preparation of sulfur hydrosols is employed to illustrate condensation or aggregation methods. Sulfur is insoluble in water but somewhat soluble in alcohol. When an alcoholic solution of sulfur is mixed with water, a bluish white colloidal dispersion results. In the absence of added stabilizing agents, the particles tend to agglomerate and precipitate on standing. This technique of dissolving the material in a water-miscible solvent such as alcohol or acetone and producing a hydrosol by precipitation with water is applicable to many organic compounds, and has been used to prepare hydrosols of natural resins like mastic, of stearic acid and of polymers (the so-called pseudolatexes).

For sulfur, another less common physical method is to introduce a current of sulfur vapor into water. Condensation produces colloidal particles. Alternatively, the very fine powder produced by condensing sulfur vapor on cold

solid surfaces (sublimed sulfur or flowers of sulfur) can be dispersed in water by addition of a suitable surfactant to produce a hydrosol.

Chemical methods include the reaction between hydrogen sulfide and sulfur dioxide, eg, by bubbling H_2S into an aqueous SO_2 solution:

$$2\,H_2S + SO_2 \rightarrow 3\,S + 2\,H_2O$$

The same reaction occurs when aqueous solutions containing sodium sulfide and sulfite are acidified with an excess of sulfuric or hydrochloric acid. Another reaction is the decomposition of sodium thiosulfate by sulfuric acid, using either very dilute or very concentrated solutions to obtain colloidally dispersed sulfur:

$$H_2SO_4 + 3\,Na_2S_2O_3 \rightarrow 4\,S + 3\,Na_2SO_4 + H_2O$$

Both reactions also produce pentathionic acid, $H_2S_5O_6$, as a by-product. The preferential adsorption of the pentathionate anion at the surface of the sulfur particles confers a negative electric charge on the particles, stabilizing the sol (see below).[22,26,27] When powdered sulfur is boiled with a slurry of lime, it dissolves with the formation of calcium pentasulfide and thiosulfate. Subsequent acidification produces the colloidal "milk of sulfur," which on washing and drying yields Precipitated Sulfur USP (see Chapter 82).

Sols of ferric, aluminum, chromic, stannic and titanium hydroxides or hydrous oxides are produced by hydrolysis of the corresponding chlorides or nitrates:

$$AlCl_3 + 3\,H_2O \rightleftharpoons Al(OH)_3 + 3\,HCl$$

Hydrolysis is promoted by boiling the solution and/or by adding a base to neutralize the acid formed.

Double decompositions producing insoluble salts can lead to colloidal dispersions. Examples are silver chloride and nickel sulfide:

$$NaCl + AgNO_3 \rightarrow AgCl + NaNO_3$$

$$(NH_4)_2S + NiCl_2 \rightarrow NiS + 2\,NH_4Cl$$

Compare also the preparation of White Lotion, which contains precipitated zinc sulfide and sulfur (Chapter 63). Reducing salts of gold, silver, copper, mercury, platinum, rhodium and palladium with formaldehyde, hydrazine, hydroxylamine, hydroquinone or stannous chloride produces hydrosols of the metals. These are strongly colored, eg, red or blue.[21,22,27]

Radioactive Colloids—Colloidal dispersions containing radioactive isotopes find increasing diagnostic and therapeutic application in nuclear medicine. Radioactive colloids that accumulate in tumors and/or lesions or emboli, indicating their location and size, may be used as diagnostic aids. Radioactive colloids with a particle size of about 300 Å, injected intravenously, locate mainly in the reticuloendothelial systems of liver, spleen and other organs and are used in scintillation imaging. The radiation emitted by the colloids is made visible by stationary or scanning devices which show the location, size and shape of the organ being investigated, as well as any tumors within. Radiocolloids are useful in anticancer radiation therapy because of their low solubility, radiation characteristics, and their ability to accumulate and remain located in certain target organs or tumors.[30]

Colloidal gold Au 198 is made by reducing a solution of gold (^{198}Au) chloride either by treatment with ascorbic acid or by heating with an alkaline glucose solution. Gelatin is added as a protective colloid (see below). The particle size ranges from 50 to 500 Å with a mean of 300 Å. The color of the sol is cherry-red in transmitted light. Violet or blue sols have excessively large particle sizes and should be discarded. Colloidal gold is used as a diagnostic and therapeutic aid (see Chapter 33). The half-life of ^{198}Au is 2.7 days.

Technetium 99m sulfur colloid is prepared by reducing sodium pertechnetate ^{99m}Tc with sodium thiosulfate. The product, a mixture of technetium sulfide and sulfur in the colloidal particle size range, is stabilized with gelatin. It is used chiefly in liver, spleen and bone scanning. Its half-life is 6.0 hour.

Microspheres of gelatin or human serum albumin can be prepared in fairly narrow particle-size ranges from 100–200 Å through 45–55 μm. A variety of β- and γ-emitting radionuclides such as ^{131}I, ^{99m}Tc, ^{113m}In or ^{51}Cr can be incorporated to label the microspheres. Such products have been used to scan heart, brain, urogenital and gastrointestinal tracts, liver, and in pulmonary perfusion and inhalation studies.[30]

Refer to Chapters 32 and 33 for an in-depth discussion of radioisotopes.

Organic compounds that are weak bases, such as alkaloids, are usually much more soluble at lower pH values where they are ionized than at higher pH values where they exist as the free base. Increasing the pH of their aqueous solutions well above their pKa may cause precipitation of the free base. Organic compounds which are weak acids, such as barbiturates, are usually much more soluble at higher pH values where they are ionized than at lower pH values where they are in the un-ionized acid form. Lowering the pH of their solutions well below their pKa may cause precipitation of the un-ionized acid. Depending on the supersaturation of the un-ionized acids or bases and on the presence of stabilizing agents, the resultant dispersions may be in the colloidal range.

Kinetics of Particle Formation—When the solubility of a compound in water is exceeded, its solution becomes supersaturated and the compound may precipitate or crystallize. The rate of precipitation, the particle size (whether colloidal or coarse), and the particle size uniformity or distribution (whether a narrow distribution and nearly monodisperse or homodisperse particles, or a broad distribution and polydisperse or heterodisperse particles) depend on two successive and largely independent processes, nucleation and growth of nuclei.

When a solution of a salt or of sucrose is supercooled, or when a chemical reaction produces a salt in a concentration exceeding its solubility product, separation of the excess solid from the supersaturated solution is far from instantaneous. Clusters of ions or molecules called nuclei must exceed a critical size before they become stable and capable of growing into colloidal size crystals. These embryonic particles have much more surface for a given weight of material than large and stable crystals, resulting in higher surface free energy and greater solubility.

Whether *nucleation* takes place depends on the *relative supersaturation*. If C is the actual concentration of the solute before crystallization has set in, and C_s is its solubility limit, $C - C_s$ is the supersaturation and $(C - C_s)/C_s$ is the relative supersaturation. Von Weimarn recognized that the rate or velocity of nucleation (number of nuclei formed per liter per second) is proportional to the relative supersaturation. Nucleation seldom occurs at relative supersaturations below 3. The foregoing statement refers to homogeneous nucleation, where the nuclei are clusters of the same chemical composition as the crystallizing phase. If the solution contains solid impurities, such as dust particles in suspension, these may act as nuclei or centers of crystallization (heterogeneous nucleation).

Once nuclei have formed, the second process, *crystallization*, begins. Nuclei grow by accretion of ions or molecules from solution forming colloidal or coarser particles until the supersaturation is relieved, ie, until $C = C_s$. The rate of

crystallization or growth of nuclei is proportional to the supersaturation. The appropriate equation,

$$\frac{dm}{dt} = \frac{A_{sp}D}{\delta}(C - C_s)$$

is similar to the Noyes-Whitney equation governing the dissolution of particles (see Chapter 31) except that $C < C_s$ for the latter process, making dm/dt negative. In both equations, m is the mass of material crystallizing out in time t, D is the diffusion coefficient of the molecules or ions of the solute, δ is the length of the diffusion path or the thickness of the liquid layer adhering to the growing particles, and A_{sp} is their specific surface area. The presence of dissolved impurities may affect the rate of crystallization and even change the crystal habit, provided that these impurities are surface-active and become adsorbed on the nuclei or growing crystals.[22,23,25-28] For instance, 0.005% polysorbate 80 or octoxynol 9 significantly retard the growth of methylprednisolone crystals in aqueous media. Gelatin or povidone, at concentrations <0.10%, retard the crystal growth of sulfathiazole in water.

Von Weimarn found that the particle size of the crystals depends strongly on the concentration of the precipitating substance. At a very low concentration and slight relative supersaturation, diffusion is quite slow because the concentration gradient is very small. Sufficient nuclei will usually form to relieve the slight supersaturation locally. Crystal growth is limited by the small amount of excess dissolved material available to each particle. Hence, the particles cannot grow beyond colloidal dimensions. This condition is represented by points A, D and G of the schematic plot of von Weimarn (Fig 19-20). At intermediate concentrations, the extent of nucleation is somewhat greater but much more material is available for crystal growth. Coarse crystals rather than colloidal particles result (points B, E or H).

At high concentrations, nuclei appear so quickly and in such large numbers that supersaturation is relieved almost immediately, before appreciable diffusion occurs. The high viscosity of the medium also slows down diffusion of excess dissolved ions or molecules, retarding crystal growth without substantially affecting the rate of nucleation. A large number of very small particles results which, because of their proximity, tend to link, producing a translucent gel (points C and F). On subsequent dilution with water, such gels usually yield colloidal dispersions.

Thus, colloidal systems are usually produced at very low and high supersaturations. Intermediate values of supersaturation tend to produce coarse crystals. Low solubility is a necessary condition for producing colloidal dispersions. If the solubility of the precipitate is increased, for instance by heating the dispersion, a new family of curves will result, similar in shape to ABC, DEF, and GHI of Fig 19-20, but displaced upwards (towards larger particle sizes) and to the right (towards higher concentrations).[25-28]

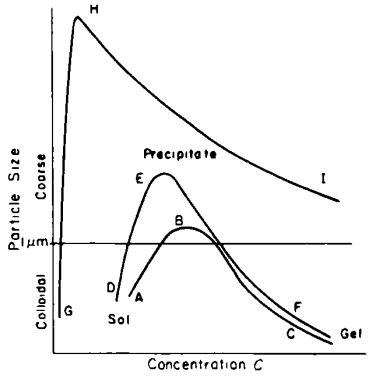


Fig 19-20. Effect of the concentration of the precipitating material and of aging on particle size.[28] Curves ABC, DEF and GHI correspond to increasing aging. Both axes are on a logarithmic scale.

Condensation methods generally produce polydisperse sols because nucleation continues while established nuclei grow. The particles in the resultant dispersion grew from nuclei formed at different times and had different growth periods.

A useful technique for preparing monodispersed sols in the colloidal range by precipitation consists in forming all the nuclei in a single, brief burst: When, in the course of the precipitation process, the rate of homogeneous nucleation becomes appreciable, a brief period of nucleation relieves the supersaturation partially to such an extent that no new nuclei form subsequently. By controlling the precipitation process, it is rendered so slow that the supersaturation remains too small for further nucleation. Therefore, the nuclei formed in the initial burst grow uniformly by diffusion of the precipitating material as the precipitation process proceeds slowly. Throughout the rest of the precipitation, the supersaturation never again reaches sufficiently high values for forming new nuclei. It is relieved by continuous growth of the existing nuclei.[23,25,31]

Controlled hydrolysis of salts of di- and trivalent cations in aqueous solution at elevated temperatures has been used to produce colloidal dispersions of metal (hydrous) oxides of uniform size and shape, in a variety of well-defined shapes (eg, sphere, lath, cube, disc, hexagonal). Complexation of the cations, concentration and temperature control the rate of hydrolysis and, hence, the chemical composition, crystallinity, shape and size of the dispersed phase.[32]

A feature of Fig 19-20 is that aging increases the particle size. Curves ABC, DEF and GHI correspond to increasing times after mixing the reagents. Typical ages are 10–30 min, several hours, and weeks or years, respectively. This gradual increase in particle size of crystals in their mother liquor is a recrystallization process called *Ostwald ripening*. Very small particles have a higher solubility than large particles of the same substance owing to their greater specific surface area and higher surface free energy. In a saturated solution containing precipitated particles of the solute in a wide range of particle sizes, the very smallest particles dissolve spontaneously and the material deposits onto the large particles. The growth of the large crystals at the expense of the very small ones occurs because this process lowers the free energy of the dispersion. As mentioned above, the most stable system is the suspension of a few coarse crystals, whereas the colloidal dispersion of a great many fine particles of the same substance is intrinsically less stable.

The spontaneous coarsening of colloidal dispersions on aging is accelerated by a relatively high solubility of the precipitate and can be retarded by lowering the solubility or by adding traces of surface-active compounds which are adsorbed at the particle surface. For instance, barium sulfate precipitated by mixing concentrated solutions of sodium sulfate and barium chloride is largely in the colloidal range and passes through filter paper. The colloidal particles gradually grow in size by Ostwald ripening, forming large crystals which can be removed quantitatively by filtration. Heating the aqueous dispersion speeds up this recrystallization by increasing the solubility of barium sulfate in water. The addition of ethyl alcohol lowers the solubility, retarding Ostwald ripening so that the dispersion remains in the colloidal state for years.

Mathematically the effect of particle size on solubility is expressed as

$$S = S_\infty \exp\left(\frac{2\gamma M}{r\rho RT}\right) \qquad (34)$$

Table XIII—Effects of Particle Size on Solubility

r (μm)	S
0.01	$7\,S_\infty$
0.10	$1.12\,S_\infty$
1.0	$1.01\,S_\infty$
10	$1.001\,S_\infty$

$M = 500$; $\gamma = 30$ ergs/cm^2; $\rho = 1$

where S is the solubility of a spherical crystal of radius r, S_∞ is the solubility of an infinitely large crystal ($r = \infty$), M is the molecular weight, ρ is the density, γ is the crystal/solvent interfacial tension, R is the gas constant and T is the absolute temperature. Only approximations can be obtained with this equation because the particles are not spheres, and γ values are different for different crystal faces. Table XIII shows the magnitude of particle size effects on the solubility for reasonable values of M, γ and ρ. It is evident that with particles in the colloidal range, ie, $r \geq 1$ μm, S values become appreciably greater than that for a coarse crystal, hence the tendency for very fine particles to dissolve and for coarse crystals to grow at the expense of the former. This difference in solubility explains why difficulty is encountered in preparing and stabilizing suspensions of very fine particles of certain substances.

Two techniques are used to increase the solubility of very slightly soluble drugs and, hence, their rate of dissolution *in vivo*. Many organic compounds exist in various polymorphic modifications. For instance, corticosterone, testosterone, sulfaguanidine and pentobarbital each have four polymorphic forms, with different melting points and crystal structures. The three metastable polymorphs have higher solubilities than the stable form. Solvates of solid drugs, eg, hydrates, have different crystalline structures and either higher or lower solubilities than the anhydrous forms. Theophylline monohydrate is less soluble than the anhydrous form while succinylsulfathiazole is less soluble than its solvate with 1-pentanol. Milling and grinding organic crystals may produce significant proportions of amorphous or strained crystalline material, which has higher solubility than the original crystalline material.[33]

Another process by which particles in colloidal dispersions grow in size is by agglomeration of individual particles into aggregates. This process, called coagulation, is discussed below.

Purification of Hydrosols by Dialysis and Ultrafiltration

Many hydrosols contain low molecular-weight, water-soluble impurities. Inorganic dispersions often contain salts formed by the reaction producing the disperse phase. Salts are especially objectionable in the case of hydrophobic dispersions because they tend to coagulate such dispersions. Protein solutions often contain salts added as part of the separation procedure. The blood of patients with renal insufficiency contains excessive concentrations of urea and other low-molecular-weight metabolites and salts. These dissolved impurities of small molecular size are removed from the colloidal dispersions by means of membranes with pore openings smaller than the colloidal particles.

Membranes—Conventional filter papers are permeable to colloidal particles as well as to small solute molecules. Among the early membranes capable of retaining colloidal particles but permeable to small solute molecules were pig's bladder and parchment. Most membranes in current use consist of cellulose, cellulose nitrate prepared from collodion, cellulose acetate or synthetic polymers, and are available in a variety of shapes, gauges, and pore sizes. *Gel cellophane* is most widely used. It consists of sheets or tubes of cellulose made by extruding cellulose xanthate solutions (viscose) through slit or annular dies into a sodium bisulfate/sulfuric acid bath which decomposes the xanthate, precipitating the regenerated cellulose in a highly swollen or gel state. If the cellulose film were permitted to dry after purification and washing with water, it would crystallize and shrink excessively, losing most of its extensive micropore structure and turning somewhat brittle. The film is therefore impregnated with glycerin before drying. Glycerin remains in the film rather than evaporating like water. It reduces the shrinkage and blocks crystallization. This action prevents the collapse of the porous gel structure and plasticizes the film, keeping it flexible. A typical dialysis tube made from sausage casing swells to about twice its thickness in water and has an average pore diameter of 34 Å. While the pore structure of cellophane films used in dialysis and ultrafiltration causes retention of colloidal particles but permits the passage of small solute molecules, osmotic membranes are only permeable to water and retain small solute molecules as well as colloidal particles.

Dialysis—The colloidal dispersion is placed inside a sac made of sausage casing dipping in water. The small solute molecules diffuse out into the water while the colloidal material remains trapped inside because of its size. The rate of dialysis is increased by increasing the area of the membrane, by stirring, and by maintaining a high concentration gradient across the membrane. For the latter purpose, the water is replenished continuously or at least frequently. A membrane configuration which provides a particularly extensive transfer area for a given volume of dispersion is the hollow fiber. A typical fiber measures 175 μm inside diameter and 225 μm outside diameter. The dispersion to be dialyzed is circulated inside a bundle of parallel fibers while water is circulated outside the fibers throughout the bundle. Dialysis of the diffusing species takes place across the thin fiber wall. Dialysis is used in the laboratory to purify sols and to study binding of drugs by proteins, as well as in some manufacturing processes.

Electrodialysis—If the low-molecular-weight impurities to be removed are electrolytes, the dialysis can be speeded up by applying an electric potential to the sol which produces electrolysis. An electrodialyzer (Fig 19-21) is divided into three compartments by two dialysis membranes supported by screens. The two outer compartments, in which the two electrodes are placed, are filled with water while the sol is placed into the center compartment. Under the influence of the applied potential, the anions migrate from the sol into the anode (right) compartment while the cations migrate into the cathode compartment. Low-molecular-weight nonelectrolyte solutes diffuse into either compartment.

Colloidal particles are usually charged and therefore tend to migrate towards the membrane sealing off the compartment with the electrode of opposite charge. The combination of electrophoresis (see below) and gravitational sedimentation produces the accumulation of negatively charged sol particles shown in Fig 19-21. Hence the supernatant

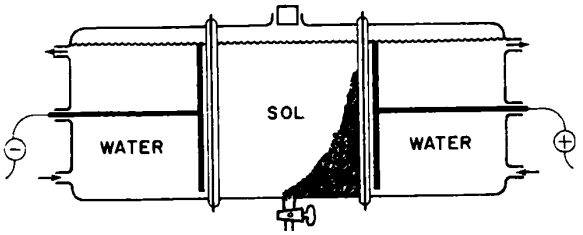


Fig 19-21. Electrodialyzer showing electrodecantation

liquid can be changed by decantation. This process, which may be used to speed up electrodialysis, is called *electrodecantation.*[21,25]

Ultrafiltration—When a sol is placed in a compartment closed by a dialysis membrane and pressure is applied, the liquid and the small solute molecules are forced through the membrane while the colloidal particles are retained. This process, called ultrafiltration, is based on a sieving mechanism in which all components smaller than the pore size of the filter membrane pass through it. The pressure difference required to push the dispersion medium through the ultrafilter is provided by gas pressure applied on the sol side or by suction on the filtrate side. The membrane is usually supported on a fine wire screen.[24-27]

As ultrafiltrate is being removed, the sol becomes more concentrated because a constant amount of disperse particles is confined to a decreasing volume of liquid. Some dissolved small molecules or ions are left in the sol together with the residual water. To avoid the increase in concentration of the colloidal particles and remove the dissolved impurities completely, the ultrafiltrate squeezed from the sol is replenished continuously or intermittently with an equal volume of water. During ultrafiltration, solids tend to accumulate on and near the membrane. To prevent this buildup and maintain uniform composition throughout the sol, it is stirred.

Bundles of hollow fibers are used for ultrafiltration in the laboratory and on large scale. To withstand higher pressures, the wall thickness of the fibers used in ultrafiltration is usually greater than that of fibers used exclusively for dialysis. When hollow fibers are fouled by excessive accumulation of solids on the inner wall, they are cleaned by backflushing with water or ultrafiltrate.

Hemodialysis—The blood of uremic patients is dialyzed periodically in "artificial kidney" dialyzers to remove urea, creatinine, uric acid, phosphate and other metabolites, and excess sodium and potassium chloride. The dialyzing fluid contains sodium, potassium, calcium, chloride and acetate ions (the latter are converted in the body to bicarbonate), dextrose and other constituents in the same concentration as normal plasma. Since it contains no urea, creatinine, uric acid, phosphate nor any of the other metabolites normally eliminated by the kidneys, these compounds diffuse from the patient's blood into the dialyzing fluid until their concentration is the same in blood and fluid. Sodium and potassium chloride diffuse from blood to fluid because of their higher initial concentration in the blood, and continue to diffuse until the concentration is equalized. The volume of dialyzing fluid is much greater than that of blood. The great disparity in volume and the replenishment of dialyzate with fresh fluid ensure that the metabolites and the excess of electrolytes are removed almost completely from the blood. Hemodialysis is also employed in acute poisoning cases.

Plasma proteins and blood cells cannot pass through the dialysis membrane because of their size. Edema resulting from water retention can be relieved by ultrafiltration through the application of a slight pressure on the blood side or a partial vacuum on the fluid side.

The three geometries used to circulate the blood and the dialyzing fluid in a countercurrent fashion are a coil of flattened cellulose tubing wound concentrically with a supporting mesh screen around a core, a stack of flat cellulose sheets separated by ridged or grooved plates, and hollow fibers. The regenerated cellulose used in the former two is precipitated from a cuprammonium solution. The hollow cellulose acetate fibers have an outside diameter of about 270 μm and a wall thickness of 30 μm.[34] The advantage of hollow fibers is their compactness. A bundle of 10,000 fibers 18 cm long has a surface area of 1.4 m^2.

Particle Shape, Optical, and Transport Properties of Lyophobic Dispersions

Hydrophobic materials handled by pharmacists in aqueous dispersion range from metallic conductors to inorganic precipitates to organic solids and liquids which are electric insulators. Despite the great diversity of the hydrophobic disperse phase, their hydrosols have certain common characteristics.

Particle Shape and Particle Size Distribution—Both of these properties depend on the chemical and physical nature of the disperse phase and on the method employed to prepare the dispersion. Primary particles exist in a great variety of shapes. Their aggregation produces an even greater variety of shapes and structures. Precipitation and mechanical comminution generally produce randomly shaped particles unless the precipitating solids possess pronounced crystallization habits or the solids being ground possess strongly developed cleavage planes. Precipitated aluminum hydroxide gels and micronized particles of sulfonamides and other organic powders have typical irregular random shapes. An exception is bismuth subnitrate. Even though its particles are precipitated by hydrolyzing bismuth nitrate solutions with sodium carbonate, its particles are lath-shaped. Precipitated silver chloride particles have a cubic habit which is apparent under the electron microscope. Lamellar or plate-like solids in which the molecular cohesion between layers is much weaker than within layers frequently preserve their lamellar shape during mechanical comminution, because milling and micronization break up stacks of thin plates in addition to fragmenting plates in the lateral dimensions. Examples are graphite, mica and kaolin. Figure 19-22 shows a Georgia crude clay as mined. Processing yields the refined, fine-particle kaolinite of Fig 19-23. Similarly, macroscopic asbestos and cellulose fibers consist of bundles of microscopic and submicroscopic fibrils. Mechanical comminution or beating splits these bundles into the component fibrils of very small diameters as well as cutting them shorter.

Microcrystalline cellulose is a fibrous thickening agent and tablet additive made by selective hydrolysis of cellulose.

Fig 19-22. Scanning electron micrograph of a crude kaolin clay as mined. Processing yields the fine particle material of Fig 19-23 (courtesy, John L Brown, Engineering Experiment Station, Georgia Institute of Technology).

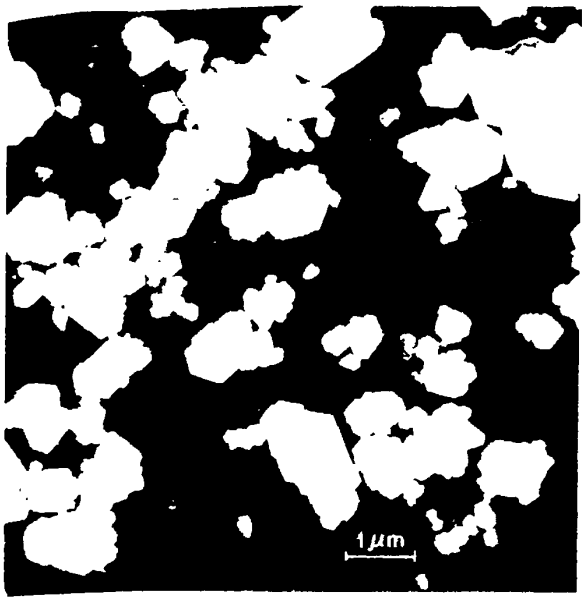


Fig 19-23. Transmission electron micrograph of a well crystallized, fine-particle kaolin. Note hexagonal shape of the clay platelets (courtesy, John L Brown, Engineering Experiment Station, Georgia Institute of Technology).

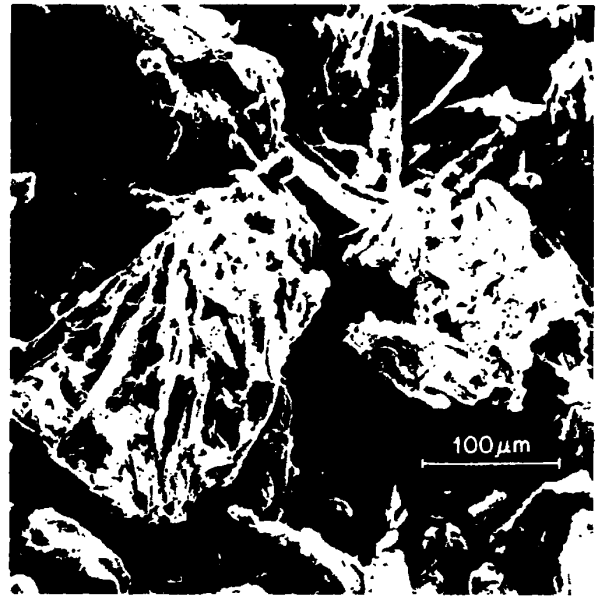


Fig 19-24. Scanning electron micrograph of Avicel PH-102 tableting grade microcrystalline cellulose. The aggregates of fiber bundles are porous and compressible (courtesy, FMC Corporation; Avicel is a registered trademark of FMC Corporation).

Native cellulose consists of crystalline regions where the polymer chains are well aligned and in registry, with maximum interchain attraction by secondary valence forces, called crystallites, and of more disordered regions having lower density and reduced interchain attraction and crystallinity, the so-called "amorphous" regions. During treatment with dilute mineral acid, the acid penetrates the amorphous regions relatively fast and hydrolyzes the polymer chains into water-soluble fragments. If the acid is washed out before it penetrates the crystalline regions appreciably, the crystallites remain intact. Wet milling and spray-drying the aqueous suspension produces spongy and porous aggregates of rod-shaped or fibrillar bundles shown in Fig 19-24. These aggregates, averaging 100 μm in size, were embrittled by the acid treatment and lost the elasticity of the native cellulose. They are well compressible and capable of undergoing plastic deformation, a property important in tableting. Their porosity permits the aggregates to absorb liquid ingredients while still remaining a free-flowing powder, thus preventing these liquids from reducing the flowability of the granulation or direct-compression mass during tableting. The swelling of the cellulosic particles in water speeds up the disintegration of the ingested tablets.

Additional shear breaks up the aggregated bundles into the individual, needle- or rod-shaped cellulose crystallites shown in Fig 19-25. The latter, which average 0.3 μm in length and 0.02 μm in width, are of colloidal dimensions. These primary particles act as suspending agents in water, producing thixotropic structured vehicles. At concentrations above 10%, eg 14 or 15%, the cellulose microcrystals gel water to ointment consistency by swelling and producing a continuous network of rods extending throughout the entire vehicle. Attraction between the elongated particles is presumably due to flocculation in the secondary minimum (see below). Treatment of the microcrystalline mass with sodium carboxymethylcellulose facilitates its disintegration into the primary needle-shaped particles and enhances their thickening action.

While in the special cases of certain clays and cellulose, comminution produces lamellar and fibrillar particles, respectively, as a rule regular particle shapes are produced by

Fig 19-25. Transmission electron micrograph of Avicel RC-591 thickening grade microcrystalline cellulose. The needles are individual cellulose crystallites; some are aggregated into bundles (courtesy, FMC Corporation; Avicel is a registered trademark of FMC Corporation).

condensation rather than by disintegration methods. *Colloidal silicon dioxide* is called fumed or pyrogenic silica because it is manufactured by high-temperature, vapor-phase hydrolysis of silicon tetrachloride in an oxy-hydrogen flame, ie, a flame produced by burning hydrogen in a stream of oxygen. The resultant white powder consists of submicroscopic spherical particles of rather uniform size (narrow particle size distribution). Different grades are produced by different reaction conditions. Relatively large, single

spherical particles are shown in Fig 19-26. Their average diameter is 50 nm (500 Å), corresponding to the comparatively small specific surface area of 50 m^2/g. Smaller spherical particles have correspondingly larger specific surface areas; the grade with the smallest average diameter, 5 nm, has a specific surface area of 380 m^2/g. During the manufacturing process, the finer-grade particles tend to sinter or grow together into chain-like aggregates resembling pearl necklaces or streptococci (see Fig 19-27).

Since fumed silica is amorphous, its inhaled dust causes no silicosis. The spheres of colloidal silicon dioxide are nonporous. While the density of the spherical particles is 2.13 g/cm^3, the bulk density of their powder is a mere 0.05 g/cm^3; the powder is extremely light. This results in two pharmaceutical and cosmetic applications for colloidal silicon dioxide. It is used to increase the fluffiness or bulk volume of powders. Even more than microcrystalline cellulose, the high porosity of silica enables it to absorb a variety of liquids from fluid fragrances to viscous tars, transforming them into free-flowing powders that can be incorporated into tablets or capsules. The porosity in colloidal silicon dioxide is due entirely to the enormous void space between the particles, which themselves are solid.

When these ultrafine particles are incorporated at levels as low as 0.1 to 0.5% into a powder consisting of coarse particles or granules, they coat the surface of the latter and act as tiny ball bearings and spacers, improving the flowability of the powder and eliminating caking. This action is important in tableting. Moreover, colloidal silicon dioxide improves tablet disintegration.

The surface of the particles contains siloxane (Si—O—Si) and silanol (Si—OH) groups. When colloidal silicon dioxide powder is dispersed in nonpolar liquids, the particles tend to adhere to one another by hydrogen bonds between their surface groups. With finer grades of colloidal silicon dioxide, the spherical particles are linked together into short chain-like aggregates as shown in Fig 19-27, thus agglomerating into loose three-dimensional networks which increase the viscosity of the liquid vehicles very effectively at levels as low as a few percent. These hydrogen-bonded structures are torn apart by stirring but rebuilt while at rest, conferring thixotropy to the thickened liquids.

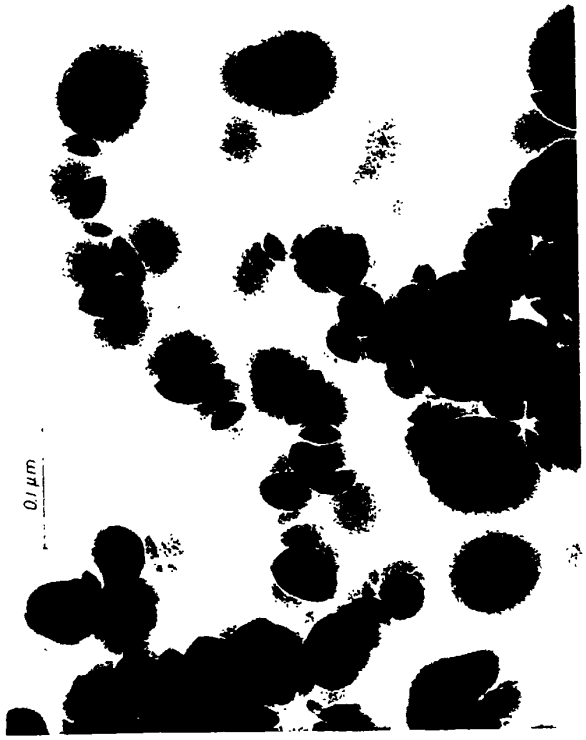


Fig 19-26. Transmission electron micrograph of Aerosil OX 50, ground and dusted on. The spheres are translucent to the electron beam, causing overlapping portions to be darker owing to increased thickness (courtesy, Degussa AG of Hanau, West Germany; Aerosil is a registered trademark of Degussa). The suffix 50 indicates the specific surface area in m^2/g.

Fig 19-27. Transmission electron micrograph of Aerosil 130, ground and dusted on. The spheres are fused together into chain-like aggregates (courtesy, Degussa AG of Hanau, West Germany; Aerosil is a registered trademark of Degussa). The suffix 130 gives the specific surface area in m^2/g.

The grades which consist of relatively large and unattached spherical particles, such as those of Fig 19-26, are less efficient thickening agents as they lack the high specific surface area and the asymmetry of the finer grades, which consist of short chains of fused spherical particles. In the latter category is Aerosil 200, the grade most widely used as a pharmaceutical adjuvant, whose primary spheres, which are extensively sintered together, have an average diameter of 12 nm. At levels of 8 to 10%, it thickens liquids of low polarity such as vegetable and mineral oils to the consistency of ointments, imparting considerable yield values to them. The consistency of ointments thickened with colloidal silicon dioxide is not appreciably reduced at higher temperatures. Incorporation of colloidal silicon dioxide into ointments and pastes, such as those of zinc oxide, also reduces the syneresis or *bleeding* of the liquid vehicles.

Hydrogen-bonding liquids like alcohols and water solvate the silica spheres, reducing the hydrogen bonding between particles. These solvents are gelled at silica levels of 12–18% or higher.

Latexes of polymers are aqueous dispersions prepared by emulsion polymerization. Their particles are spherical because polymerization of solubilized liquid monomer takes

place inside spherical surfactant micelles which swell because additional monomer keeps diffusing into the micelles. Examples include latex-based paints. Some *clays* grow as plate-like particles possessing straight edges and hexagonal angles, eg bentonite and kaolin (see Fig 19-23). Other clays have lath-shaped (nontronite) or needle-shaped particles (attapulgite).

Emulsification produces spherical droplets to minimize the oil-water interfacial area. Cooling the emulsion below the melting point of the disperse phase freezes it in the spherical shape. For instance, paraffin can be emulsified in 80° water; cooling to room temperature produces a hydrosol with spherical particles.

Sols of viruses and globular proteins, which are hydrophilic, contain compact particles possessing definite geometric shapes. Poliomyelitis virus is spherical, tobacco mosaic virus is rod-shaped, while serum albumin and the serum globulins are prolate ellipsoids of revolution (football-shaped).

Dispersion methods produce sols with wide particle size distributions. Condensation methods *may* produce essentially monodisperse sols provided specialized techniques are employed. Monodisperse polystyrene latexes are available for calibration of electron micrographs (see Fig 19-23). Biologic hydrophilic polymers, such as nucleic acids and proteins, form largely monodisperse particles, as do more highly organized structures such as lipoproteins and viruses.

Light-Scattering by Colloidal Particles—The optical properties of a medium are determined by its refractive index. When the refractive index is uniform throughout, light will pass the medium undeflected. Whenever there are discrete variations in the refractive index caused by the presence of particles or by small-scale density fluctuations, part of the light will be scattered in all directions. An optical property characteristic of colloidal systems, called the *Tyndall beam*, is familiar to everyone in the case of aerosols. When a narrow beam of sunlight is admitted through a small hole into a darkened room, the presence of the minute dust particles suspended in air is revealed by bright flashing points.

A beam of light striking a particle polarizes the atoms and molecules of that particle, inducing dipoles which act as secondary sources and reemit weak light of the same wavelength as the incident light. This phenomenon is called *light-scattering*. The scattered radiation propagates in all directions away from the particle. In a bright room, the light scattered by the dust particles is too weak to be noticeable.

Colloidal particles suspended in a liquid also scatter light. When an intense, narrowly defined beam of light is passed through a suspension, its path becomes visible because of the scattering of light by the particles in the beam. This Tyndall beam becomes most visible when viewed against a dark background in a direction perpendicular to the incident beam. The magnitude of the turbidity or opalescence depends on the nature, size and concentration of the particles. When clear mineral oil is dispersed in an equal volume of a clear aqueous surfactant solution, the resultant emulsion is milky white and opaque due to light scattering. Microemulsions, where the emulsified droplets are about 40 nm (400 Å) in diameter, ie, much smaller than the wavelength of visible light, are transparent and clear to the naked eye.

The *dark-field microscope* or *ultramicroscope*, which permits observation of particles much smaller than the wavelength of light, was the only means of detecting submicroscopic particles before the advent of electron microscopy. A special cardioid condenser produces a hollow cylinder of light and converges it into a hollow cone focused on the sample. The sample is at the apex of the cone, where the light intensity is high. After passing through the sample, the cone of light diverges and passes outside of the microscope objective. A homogeneous sample thus gives a dark field. A similar effect can be produced with a regular Abbe condenser outfitted with a central stop and a strong light source. Colloidal particles scatter light in all directions. Some of the scattered light enters the objective and shows up the particles as bright spots. Thus, even particles smaller than the wavelength of light can be detected, provided their refractive index differs sufficiently from that of the medium. Dissolved polymer molecules and highly solvated gel particles do not scatter enough light to become visible. Asymmetric particles like flat bentonite platelets give flashing effects as they rotate in Brownian motion, because they scatter more light with their basal plane perpendicular to the light beam than edgewise. Brownian motion, sedimentation, electrophoretic mobility, and the progress of flocculation can be studied with the dark-field microscope. Polydispersity can be estimated qualitatively because larger particles scatter more light and appear brighter. The resolving power of the ultramicroscope is no greater than that of the ordinary light microscope. Particles closer together than 0.2 μm appear as a single blur.

Turbidity may be used to measure the concentration of dispersed particles in two ways. In *turbidimetry*, a spectrophotometer or photoelectric colorimeter is used to measure the intensity of the light transmitted in the incident direction. Turbidity, τ, is defined by an equation analogous to Beer's law for the absorption of light (see Chapter 30),[24,25,27] namely

$$\tau = \frac{1}{l} \ln \frac{I_0}{I_t}$$

where I_0 and I_t are the intensities of the incident and transmitted light beams, and l is the length of the dispersion through which the light passes.

If the dispersion is less turbid, the intensity of light scattered at 90° to the incident beam is measured with a *nephelometer*. Both methods require careful standardization with suspensions containing known amounts of particles similar to those to be measured. The concentration of colloidal dispersions of inorganic and organic compounds and of bacterial suspensions can thus be measured by their turbidity.

The turbidity or Tyndall effect of hydrophilic colloidal systems like aqueous solutions of gums, proteins and other polymers is far weaker than that of lyophobic dispersions. These solutions appear clear to the naked eye. Their turbidity can be measured with a photoelectric cell/photomultiplier tube and serves to determine the molecular weight of the solute.

The theory of light scattering was developed in detail by Lord Rayleigh. For white nonabsorbing nonconductors or dielectrics like sulfur and insoluble organic compounds, the equation obtained for spherical particles whose radius is small compared to the wavelength of light λ is[24-27]

$$I_s = I_0 \frac{4\pi^2 n_0^2 (n_1 - n_0)^2}{\lambda^4 d^2 c} (1 + \cos^2 \theta)$$

I_0 is the intensity of the unpolarized incident light; I_s is the intensity of light scattered in a direction making an angle θ witb the incident beam and measured at a distance d. The scattered light is largely polarized. The concentration c is expressed as the number of particles per unit volume. The refractive indices n_1 and n_0 refer to the dispersion and the solvent, respectively.

Since the intensity of scattered light is inversely proportional to the fourth power of the wavelength, blue light ($\lambda \cong$ 450 nm or 4500 Å) is scattered much more strongly than red light ($\lambda \cong$ 650 nm or 6500 Å). With incident white light, colloidal dispersions of colorless particles appear blue when

viewed in scattered light, ie, in lateral directions such as 90° to the incident beam. Loss of the blue rays due to preferential scattering leaves the transmitted light yellow or red. Preferential scattering of blue radiation sideways accounts for the blue color of the sky, sea, cigarette smoke, and diluted milk and for the yellow-red color of the rising and setting sun viewed head-on.

The particles in pharmaceutical suspensions, emulsions and lotions are generally larger than the wavelength of light λ. When the particle size exceeds λ/20, destructive interference between light scattered by different portions of the same particle lowers the intensity of scattered light and changes its angular dependence. Rayleigh's theory was extended to large and to strongly absorbing and conducting particles by Mie and to nonspherical particles by Gans.[21,22,24-27] By using appropriate precautions in experimental techniques and in interpretation, it is possible to determine an average particle size and even the particle size distribution of colloidal dispersions and coarser suspensions by means of turbidity measurements.

Diffusion and Sedimentation—The molecules of a gas or liquid are engaged in a perpetual, random thermal motion which causes them to collide with one another and with the container wall billions of times per second. Each collision changes the direction and the velocity of the molecules involved. Dissolved molecules and suspended colloidal particles are continuously and randomly buffeted by the molecules of the suspending medium. This random bombardment imparts to solutes and particles an equally unceasing and erratic movement called *Brownian motion*, after the botanist Robert Brown who first observed it under the microscope with an aqueous pollen suspension. The Brownian motion of colloidal particles mirrors on a magnified scale the random movement of the molecules of the liquid or gaseous suspending medium, and represents a three-dimensional random walk.

Solute molecules and suspended colloidal particles undergo rotational and translational Brownian movement. For the latter, Einstein derived the equation

$$\bar{x} = \sqrt{2Dt}$$

where $\bar{x}$ is the mean displacement in the x-direction in time t and D is the *diffusion coefficient*. Einstein also showed that for spherical particles of radius r under conditions specified in Chapter 20 for the validity of Stokes' law and Einstein's law of viscosity

$$D = \frac{RT}{6\pi\eta rN}$$

where R is the gas constant, T the absolute temperature, N Avogadro's number, and η the viscosity of the suspending medium.

The diffusion coefficient is a measure of the mobility of a dissolved molecule or suspended particle in a liquid medium. Representative values at room temperature, in cm^2/sec, are 4.7×10^{-6} for sucrose and 6.1×10^{-7} for serum albumin in water. With a diffusion coefficient of 1×10^{-7} cm^2/sec, Brownian motion causes a particle to move by an average distance of 1 cm in one direction in 58 days, by 1 mm in 14 hr, and by 1 μm in 0.05 sec. Smaller molecules diffuse faster in a given medium. Assuming spherical shape, the radius of a serum albumin molecule is 35 Å and that of a sucrose molecule 4.4 Å. The ratio of the radii of the two molecules 35/4.4 = 7.9, is nearly identical with the inverse ratio of their diffusion coefficients in water, $4.7 \times 10^{-6}/6.1 \times 10^{-7} = 7.7$, in agreement with the above equation. Diffusion coefficients of steroids and other molecules of similar size dissolved in absorption bases based on petrolatum are generally in the 10^{-10} to 10^{-8} cm^2/sec range. Steroids have only slightly higher molecular weights than sucrose. Their much smaller diffusion coefficients are due to the much higher viscosity of the vehicle.

Dynamic light-scattering or photon-correlation spectroscopy is based on the fact that the light scattered by particles in Brownian motion undergoes a minute shift in wavelength by the usual Doppler effect. The shift is so small that it can be detected only by laser light beams, which are strictly monochromatic and very intense. The wavelength shift, which shows up as line broadening, is used to determine the diffusion coefficient of the particles,[23,26] which in turn yields their radius according to the equation above.

Brownian motion and convection currents maintain dissolved molecules and small colloidal particles in suspension indefinitely. As the particle size and r increase, the Brownian motion decreases; $\bar{x}$ is proportional to $r^{-1/2}$. Provided that the density of the particle d_P and of the liquid vehicle d_L are sufficiently different, larger particles have a greater tendency to settle out when $d_P > d_L$ or to rise to the top of the suspension when $d_P < d_L$ than smaller particles of the same material.

The rate of *sedimentation* is expressed by the Stokes' equation (Eq 35), which can be rewritten as

$$h = \frac{2(d_P - d_L)r^2gt}{9\eta}$$

where h is the height through which a spherical particle settles in time t. The rate of sedimentation is proportional to r^2. Thus, with increasing particle size, the Brownian motion diminishes while the tendency to sediment increases. The two become equal for a critical radius when the distance h through which the particle settles equals the mean displacement $\bar{x}$ due to Brownian motion in the same time interval t.[35] In most pharmaceutical suspensions, sedimentation prevails. Intravenous vegetable oil emulsions do not tend to cream because the mean droplet size, *ca* 0.5 μm, is smaller than the critical radius.

Passive diffusion caused by a concentration gradient and carried out through Brownian motion is important in the release of drugs from topical preparations (see Chapter 87) and in the gastrointestinal absorption of drugs (see Chapter 35).

Viscosity—Most lyophobic dispersions have viscosities not much greater than that of the liquid vehicle. This holds true even at comparatively high volume fractions of the disperse phase unless the particles form continuous network aggregates throughout the vehicle, in which case yield values are observed. Most O/W and W/O emulsions have specific viscosities not much greater than those predicted by Einstein's modified law of viscosity (see Eq 11 of Chapter 20 and text). For instance, emulsions containing 40% v/v of the internal phase generally have viscosities only three to five times higher than that of the continuous phase. By contrast, the apparent viscosities of lyophilic dispersions, especially of polymer solutions, are several orders of magnitude greater than the viscosity of the solvent or vehicle even at concentrations of only a few percent solids. Lyophilic dispersions are also generally much more pseudoplastic or shear-thinning than lyophobic dispersions (see Chapter 20).

Electric Properties and Stability of Lyophobic Dispersions

Difference between Lyophilic and Lyophobic Dispersions—*Lyophilic* or solvent-loving solids are called hydrophilic if the solvent is water. Owing to the presence of high concentrations of hydrophilic groups, they dissolve or disperse spontaneously in water as far as is possible without breaking covalent bonds. Among hydrophilic groups are ionized ones which dissociate into highly hydrated ions like carboxylate, sulfonate or alkylammonium ions, and organic

functional groups like hydroxyl, carbonyl, amino, and imino which bind water through hydrogen bonding.

The free energy of dissolution or dispersion, ΔG_s, of hydrophilic solids includes a large negative (exothermic) heat or enthalpy of solvation, ΔH_s, and a large increase in entropy, ΔS_s. Since $\Delta G_s = \Delta H_s - T\Delta S_s$, ΔG_s has a large negative value: the dissolution of hydrophilic macromolecules and the dispersion of hydrophilic particulate solids in water occur spontaneously (see Chapter 16), overcoming the parallel increases in surface area and surface free energy. Dissolution and dispersion take place so that water can come into contact and interact with the hydrophilic groups of the solids (enthalpy of solvation), and to increase the number of available configurations of the macromolecules and particles (entropy increase).

The van der Waals energies of attraction between dissolved macromolecules or dispersed hydrophilic solid particles are smaller than ΔG_s and are, therefore, insufficient to cause separation of a solid polymer phase or agglomeration through flocculation or coagulation of the dispersed particles. Furthermore, the hydration layer surrounding dissolved macromolecules and dispersed particles forms a barrier preventing their close approach.

Hydrophobic solids and liquids such as organic compounds consisting largely of hydrocarbon portions with few if any hydrophilic functional groups, like cholesterol and other steroids, and some nonionized inorganic substances like sulfur, are hydrated slightly or not at all. Hence they do not disperse or dissolve spontaneously in water: ΔG_s is positive because of a positive (endothermic) ΔH_s term, making the reverse process (agglomeration) the spontaneous one. Aqueous dispersions of such hydrophobic solids or liquids can be prepared by physical means which supply the appropriate energy to the system (see above). They are unstable, however. The van der Waals attractive forces between the particles cause them to aggregate, since the solvation forces which promote dispersal in water are weak. If aqueous dispersions of hydrophobic solids are to resist reaggregation (coagulation and flocculation), they must be stabilized. Stabilizing factors include electric charges at the particle surface (due to dissociation of ionogenic groups of the solid or pertaining to adsorbed ions such as ionic surfactants) and the presence of adsorbed macromolecules or nonionic surfactants. These stabilizing factors do not alter the intrinsic thermodynamic instability of lyophobic dispersions; ΔG_s is still positive so that the reverse process of phase separation or aggregation is energetically favored over dispersal. They establish kinetic barriers which delay the aggregation processes almost indefinitely; the dispersed particles cannot come together close enough for the van der Waals attractive forces to produce coagulation.[24,26,27] These stabilization mechanisms are discussed below.

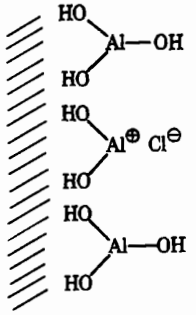
The reductions in surface area and surface free energy accompanying flocculation or coagulation are small because irregular solid particles, being rigid, touch only at a few points upon aggregation. The loose initial contacts may grow with time by sintering or recrystallization. Sintering consists of the "fusion" of primary particles into larger primary particles which propagates from initial small areas of contact. This recrystallization process is spontaneous because it decreases the specific surface area of the disperse solid and the surface free energy of the dispersion. Sintering is analogous to Ostwald ripening, the recrystallization process of transferring solid from colloidal to coarse particles discussed above. Low solubility and the presence of adsorbed surface-active substances retard both processes.

Origin of Electric Charges—Particles can acquire charges from several sources. In *proteins*, one end group of the polypeptide chain and aspartic and glutamic acid units contribute carboxylic acid groups, which are ionized into carboxylate ions in neutral to alkaline media. The other chain end group and lysine units contribute amino groups, arginine units contribute guanidine groups, and histidine units contribute imidazole groups. The nitrogen atoms of these groups become protonated in neutral to acid media. For electroneutrality, these cationic groups require anions, such as Cl^- if hydrochloric acid was used to make the medium acid and to supply the protons. The neutralizing ions, called counterions, dissociate from the ionogenic basic functional groups and can be replaced by other ions of like charge: they are not an integral part of the protein particle but are located in its immediate vicinity. The alkylammonium, guanidinium and imidazolium ions, which are attached to the protein molecule by covalent bonds, confer a positive charge to it. In neutral and alkaline media, Na^+, K^+, Ca^{2+} and Mg^{2+} are among the counterions neutralizing the negative charges of the carboxylate groups. The latter are covalently attached to and constitute an integral part of the protein particle, conferring a negative charge to it.

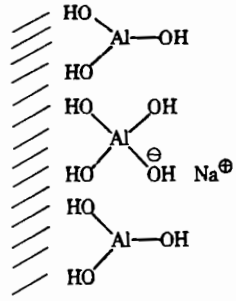
At an intermediate pH value, which ranges from 4.5 to 7 for the various proteins, the carboxylate anions and the alkylammonium, guanidinium, and imidazolium cations neutralize each other exactly. There is no need for counterions since the ionized functional groups which are an integral part of the protein molecule are in exact balance. At this pH value, called the *isoelectric point*, the protein particle or molecule is neutral; its electric charge is neither negative nor positive, but zero.[22,24,27]

Many other organic polymers contain ionic groups and are, therefore, called *polyelectrolytes* (polymeric electrolytes or salts). Natural polysaccharides of vegetable origin such as acacia, tragacanth, alginic acid and pectin contain carboxylic acid groups, which are ionized in neutral to alkaline media. Agar and carrageenan as well as the animal polysaccharides heparin and chondroitin sulfate, contain sulfuric acid hemiester groups, which are strongly acidic and ionize even in acid media. Cellulosic polyelectrolytes include *sodium carboxymethylcellulose*, while synthetic carboxylated polymers include *carbomer*, a copolymer of acrylic acid.

Aluminum hydroxide, $Al(OH)_3$, is dissolved by acids and alkalis forming aluminum ions, Al^{3+}, and aluminate ions, $[Al(OH)_4]^-$, respectively. In neutral or weakly acid media, at acid concentrations too low to cause dissolution, an aluminum hydroxide particle has some positive charges attributable to incompletely neutralized positive Al^{3+} valences. The portion of the surface of an aluminum hydroxide particle represented schematically below has one such positive charge neutralized by a Cl^- counterion:



In weakly alkaline media, at base concentrations too low to transform the aluminum hydroxide particles completely into aluminate and dissolve them, they bear some negative charges due to the presence of a few aluminate groups. The portion of the particle surface represented schematically below has one such negative group neutralized by a Na^+ counterion:



At a pH of 8.5 to 9.1,[36,37] there are neither $[Al(OH)_2]^+$ nor $[Al(OH)_4]^-$ ions in the particle surface but only neutral $Al(OH)_3$ molecules. The particles have zero charge and therefore need no counterions for charge neutralization. This pH is the isoelectric point. In the case of inorganic particulate compounds such as aluminum hydroxide, it is also called zero point of charge.

Bentonite clay is a lamellar aluminum silicate. Each lattice layer consists of a sheet of hydrated alumina sandwiched between two silica sheets. Isomorphous replacement of Al^{3+} by Mg^{2+} or of Si^{4+} by Al^{3+} confers net negative charges to the thin clay lamellas in the form of cation-exchange sites resembling silicate ions built into the lattice. The counterions producing electroneutrality are usually Na^+ (sodium bentonite) or Ca^{2+} (calcium bentonite). The zero point of charge is probably close to that of quartz, silica gel and other silicates, namely, at a pH of about 1.5 to 2.

Silver iodide sols can be prepared by the reaction

$$AgNO_3 + NaI \rightarrow AgI(s) + NaNO_3$$

In the bulk of the silver iodide particles, there is a 1:1 stoichiometric ratio of Ag^+ to I^- ions. If the reaction is carried out with an excess silver nitrate, there will be more Ag^+ than I^- ions in the surface of the particles. The particles will thus be positively charged and the counterions surrounding them will be NO_3^-. If the reaction is carried out using an exact stoichiometric 1:1 ratio of silver nitrate to sodium iodide or with an excess sodium iodide, the surface of the particles will contain an excess I^- over Ag^+ ions.[24,25,27] The particles will be negatively charged, and Na^+ will be the counterions surrounding the particles and neutralizing their charges.

An additional mechanism through which particles acquire electric charges is by the adsorption of ions,[25-27] including ionic surfactants.

Electric Double Layers—The surface layer of a silver iodide particle prepared with an excess of sodium iodide contains more I^- than Ag^+ ions, whereas its bulk contains the two ions in exactly equimolar proportion. The aqueous solution in which this particle is suspended contains relatively high concentrations of Na^+ and NO_3^-, a lower concentration of I^-, and traces of H^+, OH^- and Ag^+.

The negatively charged particle surface attracts positive ions from the solution and repels negative ions: the solution in the vicinity of the surface contains a much higher concentration of Na^+, which are the counterions, and a much lower concentration of NO_3^- ions than the bulk of the solution. A number of Na^+ ions equal to the number of excess I^- ions in the surface (ie, the number of I^- ions in the surface layer minus the number of Ag^+ ions in the surface layer) and equivalent to the net negative surface charge of a particle are pulled towards its surface. These counterions tend to stick to the surface, approaching it as closely as their hydration spheres permit (Helmholtz double layer), but the thermal agitation of the water molecules tends to disperse them throughout the solution. As a result, the layer of counterions surrounding the particle is spread out. The Na^+ concentration is highest in the immediate vicinity of the negative surface, where they form a compact layer called the Stern layer, and decreases with distance from the surface, throughout a diffuse layer called the Gouy-Chapman layer: the sharply defined negatively charged surface is surrounded by a cloud of Na^+ counterions required for electroneutrality. The combination of the two layers of oppositely charged ions constitutes an electric double layer. It is illus-

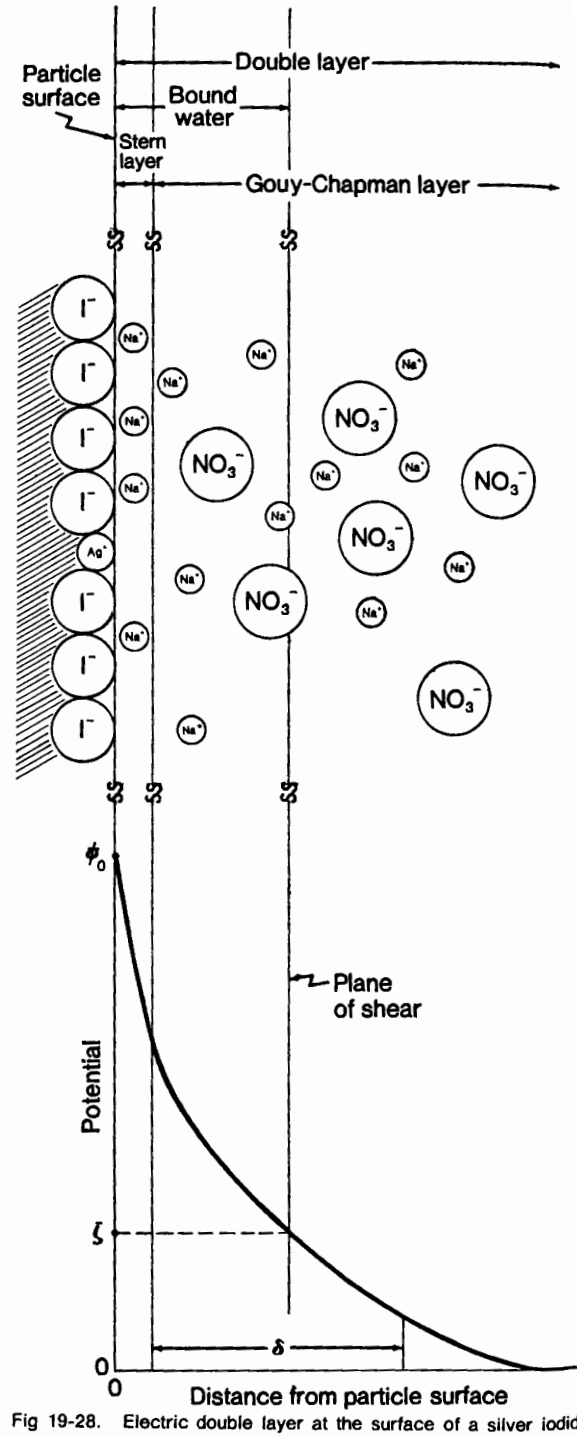


Fig 19-28. Electric double layer at the surface of a silver iodide particle (upper part) and the corresponding potentials (lower part). The distance from the particle surface, plotted on the horizontal axis, refers to both the upper and lower parts.

trated in the top part of Fig 19-28. The horizontal axis represents the distance from the particle surface in both the top and bottom parts.

The electric potential of a plane is equal to the work against electrostatic forces required to bring a unit electric charge from infinity (in this case, from the bulk of the solution) to that plane. If the plane is the surface of the particle, the potential is called surface or ψ_0 potential, which measures the total potential of the double layer. This is the thermodynamic potential which operates in galvanic cells. On moving away from the particle surface towards the bulk solution in the direction of the horizontal axis, the potential drops rapidly across the Stern layer because the Na^+ ions in the immediate vicinity of the surface screen Na^+ ions farther removed, in the diffuse part of the double layer, from the effect of the negative surface charge. The decrease in potential across the Gouy-Chapman layer is more gradual. The diffuse double layer gradually comes to an end as the composition approaches that of the bulk liquid where the anion concentration equals the cation concentration, and the potential approaches zero asymptotically. In view of the indefinite end point, the thickness δ of the diffuse double layer is arbitrarily assigned the value of the distance over which the potential at the boundary between the Stern and Gouy-Chapman layers drops to $1/e = 0.37$ of its value.[24-27] The thickness of double layers usually ranges from 10 to 1000 Å. It decreases as the concentration of electrolytes in solution increases, more rapidly for counterions of higher valence. The value of δ is approximately equal to the reciprocal of the Debye-Hückel theory parameter, κ.

Of practical importance, because it can be measured experimentally, is the electrokinetic or ζ (zeta) potential. In aqueous dispersion, even relatively hydrophobic inorganic particles and organic particles containing polar functional groups are surrounded by a layer of water of hydration attached to them by ion-dipole and dipole-dipole interaction. When a particle moves, this shell of bound water and all ions located inside it move along with the particle. Conversely, if water or a solution flows through a fixed bed of these solid particles, the hydration layer surrounding each particle remains stationary and attached to it. The electric potential at the plane of shear or slip separating the bound water from the free water is the ζ potential. It does not include the Stern layer and only that part of the Gouy-Chapman layer which lies outside the hydration shell. The various potentials are shown on the bottom part of Fig 19-28.

Stabilization by Electrostatic Repulsion—When two uncharged hydrophobic particles are in close proximity, they attract each other by van der Waals secondary valences, mainly by London dispersion forces. For individual atoms and molecules, these forces decrease with the seventh power of the distance between them. In the case of two particles, every atom of one attracts every atom of the other particle. Because the attractive forces are nearly additive, they decay much less rapidly with the interparticle distance as a result of this summation, approximately with the second or third power. Since energies of attraction are equal to force x distance, they decrease approximately with the first or second power of the distance. Therefore, whenever two particles approach each other closely, the attractive forces take over and cause them to adhere. Coagulation occurs as the primary particles aggregate into increasingly larger secondary particles or flocs.

If the dispersion consists of two kinds of particles with positive and negative charges, respectively, the electrostatic attraction between oppositely charged particles is superimposed on the attraction by van der Waals forces, and coagulation is accelerated. If the dispersion contains only one kind, as is customary, all particles have surface charges of the same sign and density. In that case, electrostatic repulsion tends to prevent the particles from approaching closely enough to come within effective range of each other's van der Waals attractive forces, thus stabilizing the dispersion against interparticle attachments or coagulation. The electrostatic repulsive energy has a range of the order of δ.

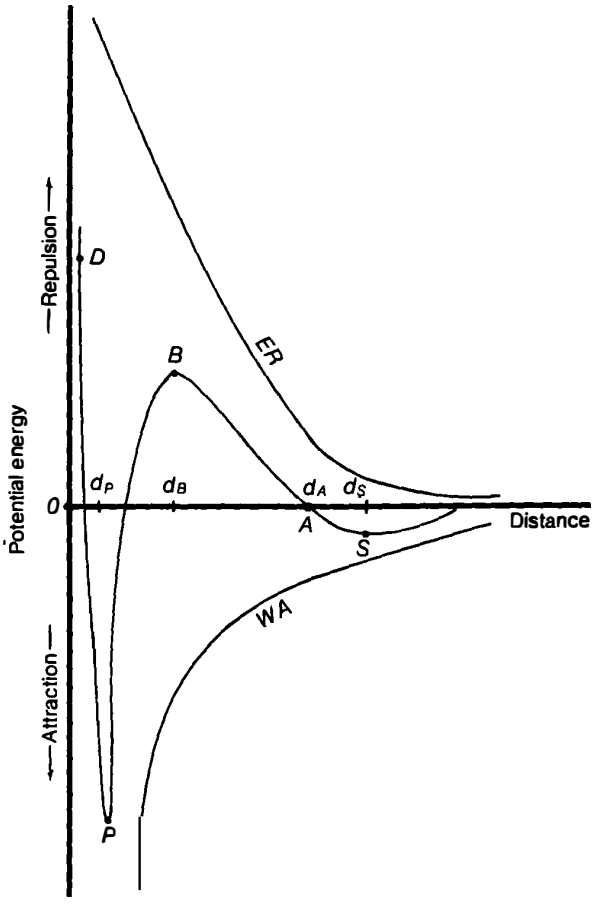


Fig 19-29. Curves representing the van der Waals energy of attraction (*WA*), the energy of electrostatic repulsion (*ER*), and the net energy of interaction (*DPBAS*) between two identical charged particles, as a function of the interparticle distance.

A quantitative theory of the interaction between lyophobic disperse particles was worked out independently by Derjaguin and Landau in the USSR and by Verwey and Overbeek in the Netherlands in the early 1940s.[21,24-27,38] Detailed calculations are also found in Chapter 21 of RPS-17. The so-called DLVO theory predicts and explains many but not all experimental data. Its refinement to account for discrepancies is still continuing.

The DLVO theory is summarized in Fig 19-29, where curve *WA* represents the van der Waals attractive energy which decreases approximately with the second power of the interparticle distance, and curve *ER* represents the electrostatic repulsive energy which decreases exponentially with distance. Because of the combination of these two opposing effects, attraction predominates at small and large distances whereas repulsion may predominate at intermediate distances. Negative energy values indicate attraction, and positive values repulsion. The resultant curve *DPBA*, obtained by algebraic addition of curves *WA* and *ER*, gives the total, net energy of interaction between two particles.

The interparticle attraction depends mainly on the chemical nature and particle size of the material to be dispersed. Once these have been selected, the attractive energy is fixed

and cannot readily be altered. The electrostatic repulsion depends on ψ_0 or the density of the surface charge and on the thickness of the double layer, both of which govern the magnitude of the ζ potential. Thus, stability correlates to some extent with this potential.[24] The ζ potential can be adjusted within wide limits by additives, especially ionic surfactants, water-miscible solvents, and electrolytes (see below). If the absolute value of the ζ potential is small, the resultant potential energy is negative and van der Waals attraction predominates over electrostatic repulsion at all distances. Such sols coagulate rapidly.

The two identical particles whose interaction is depicted in Fig 19-29 have a large (positive or negative) ζ potential resulting in an appreciable positive or repulsive potential energy at intermediate distances. They are on a collision course because of Brownian motion, convection currents, sedimentation, or because the dispersion is being stirred.

As the two particles approach each other, the two atmospheres of counterions surrounding them begin to interpenetrate or overlap at point A corresponding to the distance d_A. This produces a net repulsive (positive) energy because of the work involved in distorting the diffuse double layers and in pushing water molecules and counterions aside, which increases if the particles approach further. If the particles continue to approach each other, even after most of the intervening solution of the counterions between them has been displaced, the repulsion between their surface charges increases the net potential energy of interaction to its maximum positive value at B. If the height of the potential energy barrier B exceeds the kinetic energy of the approaching particles, they will not come any closer than the distance d_B but move away from each other. A net positive potential energy of about 25 kT units usually suffices to keep them apart, rendering the dispersion permanently stable; k is the Boltzmann constant and T is the absolute temperature. At $T = 298°K$, this corresponds to 1×10^{-12} erg. The kinetic energy of a particle is of the order of kT.

On the other hand, if their kinetic energy exceeds the potential energy barrier B, the particles continue to approach each other past d_B, where the van der Waals attraction becomes increasingly more important compared to the electrostatic repulsion. Therefore, the net potential energy of interaction decreases to zero and then becomes negative, pulling the particles still closer together. When the particles touch, at a distance d_P, the net energy has acquired the large negative value P. This deep minimum in potential energy corresponds to a very stable situation in which the particles adhere. Since it is unlikely that enough kinetic energy can be supplied to the particles or that their ζ potential can be increased sufficiently to cause them to climb out of the potential energy well P, they are attached permanently to each other. When most or all of the primary particles agglomerate into secondary particles by such a process, the sol coagulates.

Any closer approach of two particles, than the touching distance d_P, is met with a very rapid rise in potential energy along PD because the solid particles would interpenetrate each other, causing atomic orbitals to overlap (Born repulsion).

Coagulation of Hydrophobic Dispersions—The height of the potential energy barrier and the range over which the electrostatic repulsion is effective (or the thickness of the double layer) determine the stability of hydrophobic dispersions. Both factors are reduced by the addition of electrolytes. The transition between a coagulating and a stable sol is gradual and depends on the time of observation. By using standard conditions, however, it is possible to classify a sol as either coagulated or coagulating, or as stable or fully dispersed.

To determine the value of the coagulating concentration of a given electrolyte for a given sol, a series of test tubes is filled with equal portions of the sol. Identical volumes of solutions of the electrolyte, of increasing concentration, are added with vigorous stirring. After some time at rest (eg, 2 hours), the mixtures are agitated again. After an additional, shorter rest period (eg, ½ hour), they are inspected for signs of coagulation. The tubes can be classified into two groups, one showing no signs of coagulation and the other showing at least some signs, eg, visible flocs. Alternatively, they can be classified into one group showing complete coagulation and the other containing at least some deflocculated colloid left in the supernatant. In either case, the separation between the two classes is quite sharp. The intermediate agitation breaks the weakest interparticle bonds and brings small particles in contact with larger ones, thus increasing the sharpness of separation between coagulation and stability. After repeating the experiment with a narrower range of electrolyte concentrations, the coagulation value c_{CV} of the electrolyte, ie, the lowest concentration at which it coagulates the sol, is established with good reproducibility.[24,25,27]

Typical c_{CV} data for a silver iodide sol prepared with an excess of iodide are listed in Table XIV. The following conclusions can be drawn from the left half of Table XIV:

1. The c_{CV} does not depend on the valence of the anion, since nitrate and sulfate of the same metal have nearly identical values.
2. The differences among the c_{CV}s of cations with the same valence are relatively minor. However, there is a slight but significant trend of decreasing c_{CV} with increasing atomic number in the alkali and in the alkaline earth metal groups. Arranging these cations in the order of decreasing c_{CV} produces the *Hofmeister* or *lyotropic series*. It governs many other colloidal phenomena, including the effect of salts on the temperature of gelation and the swelling of aqueous gels and on the viscosity of hydrosols, the salting out of hydrophilic colloids, the cation exchange on ion-exchange resins, and the permeability of membranes toward salts. The series is also observed in many phenomena involving only small atoms or ions and true solutions, including the ionization potential and electronegativity of metals, the heats of hydration of cations, the size of the hydrated cations, the viscosity, surface tension and infrared spectra of salt solutions, and the solubility of gases therein. For monovalent cations, the lyotropic series is

$$Li^+ > Na^+ > K^+ > NH_4^+ > Rb^+ > Cs^+$$

A similar lyotropic series exists for anions.[21,22,24-26]

The lithium ion has a higher c_{CV} than the cesium ion because it is more extensively hydrated, so that Li^+ (aq), including the hydration shell, is larger than Cs^+ (aq). Owing to its smaller size, the hydrated cesium ion can approach the negative particle surface more closely than the hydrat-

Table XIV—Coagulation Values for Negative Silver Iodide Sol[a]

Electrolyte	c_{CV}, mM/L	Electrolyte	c_{CV}, mM/L
$LiNO_3$	165	$AgNO_3$	0.01
$NaNO_3$	140	½ $(C_{12}H_{25}NH_3)_2SO_4$	0.7
½ Na_2SO_4	141	Strychnine nitrate	1.7
KNO_3	136	½ Morphine sulfate	2.5
½ K_2SO_4	138		
$RbNO_3$	126		
Mean	141		
$Mg(NO_3)_2$	2.60	Quinine sulfate	0.7
$MgSO_4$	2.57		
$Ca(NO_3)_2$	2.40		
$Sr(NO_3)_2$	2.38		
$Ba(NO_3)_2$	2.26		
$Zn(NO_3)_2$	2.50		
$Pb(NO_3)_2$	2.43		
Mean	2.45		
$Al(NO_3)_3$	0.067		
$La(NO_3)_3$	0.069		
$Ce(NO_3)_3$	0.069		
Mean	0.068		

[a] From Ref 21 and unpublished data.

ed lithium ion. Moreover, because of its greater electron cloud, the Cs^+ ion is more polarizable than the Li^+ ion. Therefore, it is more strongly adsorbed in the Stern layer, which makes it a more effective coagulating agent.

3. The coagulation values depend primarily on the valence of the counterions, decreasing by one to two orders of magnitude for each increase of one in their valence (Schulze-Hardy rule). According to the DLVO theory, the coagulation values vary inversely with the sixth power of the valence of the counterions. For mono-, di- and trivalent counterions, they should be in the ratio

$$\frac{1}{1^6} : \frac{1}{2^6} : \frac{1}{3^6} \text{ or } 100 : 1.6 : 0.14$$

The mean c_{CV}'s of Table XIV are 141 : 2.45 : 0.068, or 100 : 1.7 : 0.05, in satisfactory agreement with the DLVO theory.

The following conclusion can be drawn from the right half of Table XIV:

4. The cations on the right side of Table XIV constitute obvious exceptions to the preceding. Ag^+ is the potential-determining counterion. *Potential-determining ions* are those whose concentration determines the surface potential. When silver nitrate is added to the negative silver iodide dispersion, some of its silver ions are incorporated into the negatively charged surface of the particles and lower the magnitude of their charge by reducing the excess of I^- ions in the surface. Thus, silver salts are exceptionally effective coagulating agents because they reduce the magnitude of the ψ_0 as well as of the ζ potential. Indifferent salts, which reduce only the latter, require much higher salt concentrations for comparable reductions in the ζ potential. The other potential-determining ion of silver iodide is I^-. Alkali iodides have higher c_{CV}'s than 141 millimole/liter because they supply iodide ions which enter the surface layer of the silver iodide particles and increase its excess of I^- over Ag^+ ions, thereby making ψ_0 more negative. Bromide and chloride ions act similarly but less effectively.

The principal potential-determining ion for proteins is H^+; those for aluminum hydroxide are OH^- (and hence H^+) and Al^{3+}, but also Fe^{3+} and Cr^{3+} which form mixed hydroxides with Al^{3+}.

5. The cationic surfactant in Table XIV and the alkaloidal salts, which also behave as such, constitute the second exception to the Schulze-Hardy rule. Surface-active compounds contain hydrophilic and hydrophobic moieties in the same molecule, the latter being hydrocarbon portions which by themselves are water-insoluble. Their dual nature causes these compounds to accumulate in interfaces. Dodecylammonium and alkaloidal cations displace inorganic monovalent cations from the Stern layer of a negatively charged silver iodide particle because they are attracted to it not only by electrostatic forces like sodium ions but also by van der Waals forces between their hydrocarbon moieties (dodecyl chains in the case of the dodecylammonium ions) and the solid. Because they are strongly adsorbed from solution onto the surface and do not tend to dissociate from it, surface-active cations are very effective in reducing the ζ potential of the negative silver iodide particles, ie, they have lower c_{CV} than purely inorganic cations of the same valence.

6. Anionic surfactants like those containing lauryl sulfate ions also have a tendency to be adsorbed at solid-liquid interfaces. However, because of electrostatic repulsion between the negatively charged surface of silver iodide particles whose surface layer contains an excess iodide ions and the surface-active anions, adsorption usually does not occur below the critical micelle concentration (see below). If such adsorption does occur, it increases the density of negative charges in the particle surface, raising the c_{CV} of anionic surfactants above that corresponding to their valence.

Ionic solids with surface layers containing the ionic species in near proper stoichiometric balance, and most water-insoluble organic compounds have relatively low surface charge densities. They adsorb ionic surfactants of like charge from solution even at low concentrations, which increases their surface charge densities and the magnitude of their ζ potentials, stabilizing their aqueous dispersions.

The addition of water-miscible solvents such as alcohol, glycerin, propylene glycol or polyethylene glycols to aqueous dispersions lowers the dielectric constant of the medium. This reduces the thickness of the double layer and, therefore, the range over which electrostatic repulsion is effective, and lowers the size of the potential energy barrier. Addition of solvents to aqueous dispersions tends to coagulate them. At concentrations too low to cause coagulation by themselves, solvents make the dispersions more sensitive to coagulation by added electrolytes, ie, they lower the c_{CV}.

Progressive addition of the salt of a counterion of high valence reduces the ζ potential of colloidal particles gradually to zero. Eventually, the sign of the ζ potential may be inverted and its magnitude may increase again, but in the opposite direction. The ψ_0 and ζ potentials of aqueous sulfamerazine suspensions are negative above their isoelectric points; those of bismuth subnitrate are positive. As discussed on page 297, the addition of Al^{3+} to the former and of PO_4^{3-} to the latter in large enough amounts inverts the sign of their ζ potentials; their ψ_0 potentials remain unchanged. Surface-active ions of opposite charge may also produce such charge inversion.

The superposition of the van der Waals attractive energy with its long-range effectiveness and the electrostatic repulsive energy with its intermediate-range effectiveness frequently produces a shallow minimum (designated S in Fig 19-29) in the resultant energy-distance curve at interparticle distances d_S several times greater than δ. If this minimum in potential energy is small compared to kT, Brownian motion prevents aggregation. For large particles such as those of many pharmaceutical suspensions and for particles which are large in one or two dimensions (rods and plates), the *secondary minimum* may be deep enough to trap them at distances d_S from each other. This requires a depth of several kT units. Such fairly long-range and weak attraction produces loose aggregates or flocs which can be dispersed by agitation or by removal or reduction in the concentration of flocculating electrolytes.[21,25–27,38] This reversible aggregation process involving the secondary minimum is called *flocculation*. By contrast, aggregation in the deep primary minimum P, called *coagulation*, is irreversible.

Stabilization by Adsorbed Surfactants—As discussed above, surfactants tend to accumulate at interfaces because of their amphiphilic nature. This process is an *oriented physical adsorption.* Surfactant molecules arrange themselves at the interface between water and an organic solid or liquid of low polarity in such a way that the hydrocarbon chain is in contact with the surface of the solid particle or sticks inside the oil droplet while the polar headgroup is oriented towards the water phase. This orientation removes the hydrophobic hydrocarbon chain from the bulk of the water, where it is unwelcome because it interferes with the hydrogen bonding among the water molecules, while leaving the polar headgroup in contact with water so that it can be hydrated.

Figure 19-30*A* shows schematically that at low surfactant concentration and low surface coverage, the hydrocarbon chains of the adsorbed surfactant molecules lie flat against the solid surface. At higher surfactant concentrations, the surfactant molecules are adsorbed in the upright position to permit the adsorption of more surfactant per unit surface area. Figure 19-30*B* shows a nearly close-packed monolayer of adsorbed surfactant molecules. The terminal methyl groups of their hydrocarbon tails are in contact with the hydrophobic surface and the hydrocarbon tails are in lateral contact with each other. London dispersion forces promote attraction between both types of adjoining groups. The polar headgroups protrude into the water and are hydrated.

The adsorption of ionic surfactants increases the charge density and the ζ potential of the disperse particles. These two parameters are low for organic substances lacking ionic or strongly polar groups. The increase in electrostatic repulsion among the nonpolar organic particles due to adsorption of surface-active ions stabilizes the dispersion against coagulation. This "charge stabilization" is described by the DLVO theory.

Most water-soluble nonionic surfactants are polyoxyethylated (see above): Each molecule consists of a hydrophobic hydrocarbon chain combined with a hydrophilic polyethylene glycol chain, eg $CH_3(CH_2)_{15}(OCH_2CH_2)_{10}OH$. Hydration of the 10 ether groups and of the terminal hydroxyl

group renders the surfactant molecule water-soluble. It adsorbs at the interface between a hydrophobic solid and water, with the hydrocarbon moiety adhering to the solid surface and the polyethylene glycol moiety protruding into the water, where it is hydrated. The particle surface is thus surrounded by a thin layer of hydrated polyethylene glycol chains. This hydrophilic shell forms a steric barrier which prevents close contact between particles and, hence, coagulation ("steric stabilization"). Nonionic surfactants also reduce the sensitivity of hydrophobic dispersions toward coagulation by salts, ie, they increase the coagulation values.[39]

In a flocculated dispersion, groups of several particles are agglomerated into flocs. Frequently, the particles of a floc are in physical contact. When a surfactant is added to a flocculated sol, the dissolved surfactant molecules become adsorbed at the surface of the particles. Surfactant molecules tend to pry apart flocs by wedging themselves between the particles at their areas of contact. This action opens up for surfactant adsorption additional surface area that was previously blocked by adhesion of another solid surface. The breaking up of flocs or secondary particles is defined above as deflocculation or peptization.

Ophthalmic suspensions should be deflocculated because the large particle size of flocs causes eye irritation. Parenteral suspensions should be deflocculated to prevent flocs from blocking capillary blood vessels and hypodermic syringes, and to reduce tissue irritation. Deflocculated suspensions tend to cake, however, ie, the sediment formed by gravitational settling is compact and may be hard to disperse by shaking. Caking in oral suspensions is prevented by controlled flocculation as discussed below.

Stabilization by Adsorbed Polymers—Water-soluble polymers are adsorbed at the interface between water and a hydrophobic solid if they have some hydrophobic groups that limit their water solubility and render them amphiphilic and, hence, surface-active. Such polymers also tend to accumulate at the air-water interface and lower the surface tension of the aqueous phase. A high concentration of ionic groups in polyelectrolytes tends to eliminate surface activity and the tendency to adsorb at interfaces, because the polymer is excessively water-soluble. An example is *sodium carboxymethylcellulose*. *Polyvinyl alcohol* is very water-soluble due to the high concentration of hydroxyl groups and does not adsorb extensively at interfaces. Polyvinyl alcohol is manufactured by the hydrolysis of polyvinyl acetate, which is water-insoluble. Incomplete hydrolysis of, say, only 85% of the acetyl groups produces a copolymer which is water-soluble but surface-active as well. Other surface-active polymers include methylcellulose, hydroxypropyl cellulose, high-molecular-weight polyethylene glycols (polyethylene oxides), and proteins. The surface activity of proteins is due to the presence of hydrophobic groups in the side chains at concentrations too low to cause insolubility in water. Proteins are denatured upon adsorption at air-water and solid-water interfaces.

The long, chain-like polymer molecules are adsorbed from solution onto solid surfaces in the form of loops projecting into the aqueous phase, as shown in Fig 19-31*A*, rather than lying flat against the solid substrate. Only a small portion of the chain segments of an adsorbed macromolecule is actually in contact with and adheres directly to the surface. Because of its great length, however, there are enough of such areas of contact to anchor the adsorbed macromolecule firmly onto the solid. Figure 19-30 is drawn on a much more expanded scale than Fig 19-31.

The sol particles are surrounded by a layer consisting of the adsorbed polymer chains, the water of hydration associated with them, and water trapped mechanically inside the chain loops. This sheath is an integral part of the particle surface. The layers of adsorbed polymer prevent the particles from approaching each other closely enough for the interparticle attraction by London dispersion forces to produce coagulation. These forces are effective only over very small interparticle distances of less than twice the thickness of the adsorbed polymer layer.

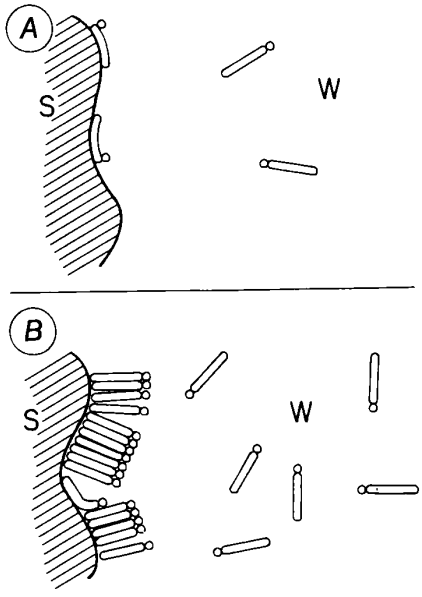


Fig 19-30. Schematic representation of the physical adsorption of surfactant molecules at a hydrophobic solid (S)/water (W) interface. Cylindrical portions and spheres represent hydrocarbon chains and polar headgroups of the surfactant molecules, respectively (*A*) low surfactant concentration/low surface coverage; (*B*) near critical micelle concentration/surface coverage near saturation

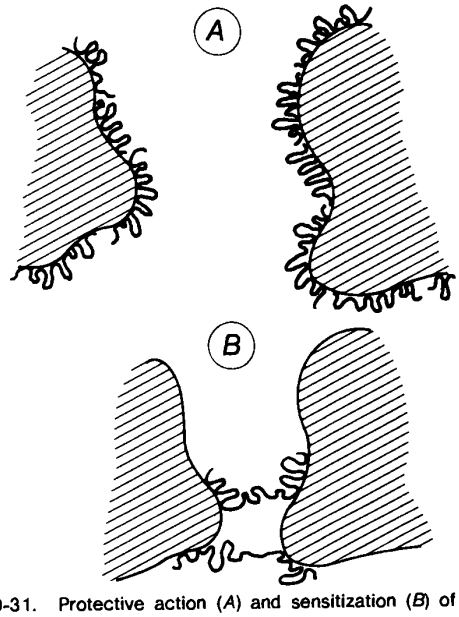


Fig 19-31. Protective action (*A*) and sensitization (*B*) of sols of hydrophobic particles by adsorbed polymer chains.

The mechanisms of *steric stabilization* by which adsorbed nonionic macromolecules prevent coagulation of hydrophobic sols (*protective action*) are also operative in the stabilization of sols by nonionic surfactants. The difference between adsorbed nonionic surfactants and adsorbed polymers

is that the hydrophilic polyethylene glycol moieties of the adsorbed surfactant molecules protruding into water resemble the chain ends of the adsorbed macromolecules rather than their looped segments. The following protective mechanisms are operative:

1. The layer of adsorbed polymer and enmeshed water surrounding the particles forms a *mechanical* or *steric barrier* between them that prevents the close interparticle approach necessary for coagulation. At dense surface coverage, these layers are somewhat elastic. They may be dented by a collision between two particles but tend to spring back.
2. When two particles approach so closely that their adsorbed polymer layers overlap, the chain loops of the two opposing layers compress and mix with or interpenetrate each other. The resulting restriction to the freedom of motion of the chain segments in the overlap region produces a negative entropy change which tends to make the free energy change for the reduction in interparticle distance required for coagulation positive. The reverse process of disentanglement of the two opposing adsorbed polymer layers resulting from separation of the particles occurs because it is energetically more favorable. The particles are thus prevented from coagulation by *entropic repulsion* through the mechanism of *entropic stabilization* of the sol. This mechanism predominates when the concentration of polymer in the adsorbed layer is low.
3. As the polymer layers adsorbed on two approaching particles overlap and compress or interpenetrate each other, more polymer segments become crowded into a given volume of the aqueous region between the particles. The increased polymer concentration in the overlap region causes a local increase in osmotic pressure, which is relieved by an influx of water. This influx to dilute the polymer loops pushes the two particles apart, preventing coagulation.
4. If the adsorbed polymer has some ionic groups, stabilization by electrostatic repulsion or charge stabilization described above is added to the three steric stabilization mechanisms to prevent a close interparticle approach and, hence, coagulation.
5. The adsorption of water-soluble polymers changes the nature of the surface of the hydrophobic particles to hydrophilic, resulting in an increased resistance of the sol to coagulation by salts.[40]

The water-soluble polymers whose adsorption stabilizes hydrophobic sols and protects them against coagulation are called *protective colloids*. *Gelatin* and *serum albumin* are the preferred protective colloids for stabilizing parenteral suspensions because of their biocompatibility. These two polymers, as well as casein (milk protein), dextrin (partially hydrolyzed starch) and vegetable gums like acacia and tragacanth are metabolized in the human body. Cellulose derivatives and most synthetic protective colloids such as *povidone* are not biotransformed. Because of this and because of their large molecular size, polymers pertaining to the last two categories are not absorbed but excreted intact when they are administered in an oral dosage form.

A semiquantitative assessment of the stabilizing efficiency of protective colloids is the *gold number*, developed by Zsigmondy. It is the largest number of milligrams of a protective colloid which, when added to 10 mL of a special standardized gold sol, just fails to prevent the change in color from red to blue on addition of 1 mL of 10% NaCl solution. The gold sol contains 0.0058% gold with a particle size of about 250 Å. Coagulation by sodium chloride causes the color change. Representative gold numbers are 0.005 to 0.01 for gelatin, 0.01 for casein, 0.02 to 0.5 for egg albumin, 0.15 to 0.5 for acacia, and 1 to 7 for dextrin.[22,27] Gelatin is a more effective protective colloid than acacia or dextrin because the presence of some hydrophobic side groups makes it more surface active and causes more extensive adsorption from solution. Other protective numbers are based on different hydrophobic disperse solids, eg, silver, Prussian blue, sulfur, ferric oxide. The ranking of different protective colloids depends somewhat on the substrate. When formulating a disperse dosage form, one should measure the protective action on the actual solid hydrophobic phase to be dispersed as a sol.

Sensitization is the opposite of protective action, namely, a decrease in the stability of hydrophobic sols. It is brought about by some protective colloids, at concentrations well below those at which they exert a protective action. A protective colloid may, at very low concentrations, flocculate a sol in the absence of added salts and/or lower the coagulation values of the sol.

In the case of nonionic polymers or of polyelectrolytes with charges of the same sign as the sol, flocculation is the result of the bridging mechanism illustrated in Fig 19-31*B*. At very low polymer concentrations, there are not nearly enough polymer molecules present to cover each sol particle completely. Since the particle surfaces are largely bare, a single macromolecule may be adsorbed on two particles, bridging the gap between them and pulling them close together. Flocs of several particles are formed when one particle is bridged or connected to two or more other particles by two or more polymer molecules adsorbed jointly on two or possibly even three particles. Such flocculation usually occurs over a narrow range and at very low values of polymer concentrations. At higher concentrations, when enough polymer is available to cover the surface of all particles completely, bridging is unlikely to occur and the adsorbed polymer stabilizes or peptizes the sol.[23,40]

The nonionic Polymer A of Fig 19-32 stabilizes the sol at all concentrations. Neither sensitization by bridging nor by charge neutralization is observed. The reason that Polymer A lowers the positive ζ potential of the sol slightly is that increasing amounts of adsorbed polymer chains gradually shift the plane of shear outward, away from the positively charged surface. If Polymer A was a cationic polyelectrolyte, the ζ potential–protective colloid concentration plot would gradually rise with increasing polymer adsorption rather than drop.

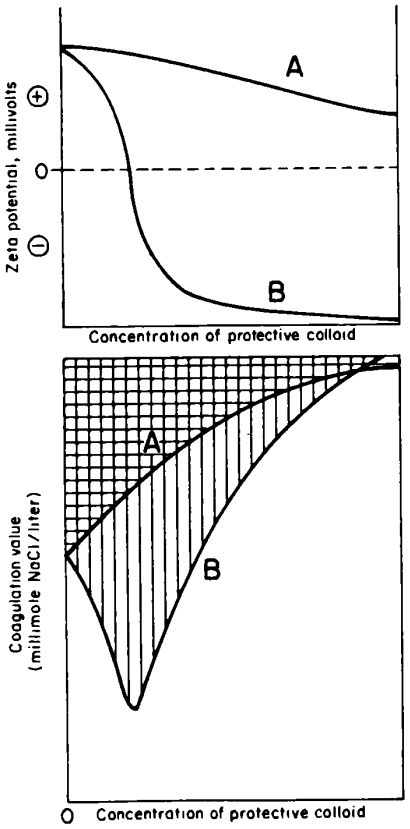


Fig 19-32. Protective action and sensitization: Polymer A exerts protective action at all concentrations, while Polymer B sensitizes at low concentrations and stabilizes at high concentrations. Horizontal and vertical hatching indicates region of flocculation for a sol treated with various concentrations of Polymers A and B, respectively. Clear region underneath indicates sol is deflocculated.

If the polymer has ionic groups of charge opposite to the charge of the sol particles, limited adsorption neutralizes the charge of the particles, reducing their ζ potential to near zero. With stabilization by electrostatic repulsion thus inoperative, and steric stabilization ineffective because of low surface coverage with adsorbed polymer, the sol either coagulates by itself or is coagulated by very small amounts of sodium chloride. At higher polymer concentrations and more extensive adsorption, charge reversal of the particles to the sign of the charge of the polyelectrolyte reactivates charge stabilization and adds steric stabilization, increasing the coagulation value of the sol well above the initial value before polymer addition.

For example, a partly hydrolyzed polyacrylamide with about 20% of ammonium acrylate repeating units is an anionic polyelectrolyte. At the ppm level, the polymer flocculates aluminum hydroxide sols at a pH of 6 to 7, where the sols are positively charged and the polyelectrolyte is fully ionized. At a polymer concentration of 1:10,000, the sol becomes negatively charged because extensive polymer adsorption introduces an excess of $—COO^-$ groups over $=Al^+$ ions into the particle surface. Steric stabilization plus electrostatic repulsion make the sol more stable against flocculation by salts than it was before the polyacrylamide addition.

Polymer B in Fig 19-32 illustrates this example. The curve in the lower plot indicates sensitization, with the coagulation value of sodium chloride lowered by as much as 60%. Zeta potential measurements can distinguish between sensitization by bridging and by charge neutralization. The charge reversal caused by adsorption of Polymer B shown in the upper plot pinpoints charge neutralization as the cause of sensitization. If Polymer B had a ζ potential–polymer concentration plot similar to Polymer A, sensitization would be ascribed to bridging.

Even water-soluble polymers which are too thoroughly hydrophilic to be adsorbed by hydrophobic sol particles can stabilize those sols. Their thickening action slows down Brownian motion and sedimentation, giving the particles less opportunity to come into contact and hence retarding flocculation.

Electrokinetic Phenomena—When a dc electric field is applied to a dispersion, the particles move towards the electrode of charge opposite to that of their surface. The counterions located inside their hydration shell are dragged along while the counterions in the diffuse double layer outside the plane of slip, in the free or mobile solvent, move toward the other electrode. This phenomenon is called *electrophoresis*. If the charged surface is immobile, as is the case with a packed bed of particles or a tube filled with water, application of an electric field causes the counterions in the free water to move towards the opposite electrode, dragging solvent with them. This flow of liquid is called *electroosmosis*, and the pressure produced by it, *electroosmotic pressure*. Conversely, if the liquid is made to flow past charged surfaces by applying hydrostatic pressure, the displacement of the counterions in the free water produces a potential difference between the two ends of the tube or bed called *streaming potential*.

The three phenomena depend on the relative motion of a charged surface and of the diffuse double layer outside the plane of slip surrounding that surface. The major part of the diffuse double layer is within the free solvent and can, therefore, move along the surface.[24—27,41] All three electrokinetic phenomena measure the identical ζ potential, which is the potential at the plane of slip.

The particles of pharmaceutical suspensions and emulsions are visible in the microscope or ultramicroscope, as are bacteria, erythrocytes and other isolated cells, latex particles, and many contaminant particles in pharmaceutical solutions. Their ζ potential is conveniently measured by *microelectrophoresis*. A potential difference E applied between two electrodes dipping into the dispersion and separated by a distance d produces the potential gradient or field strength E/d, expressed in v/cm. From the average velocity v of the particles, measured with the eyepiece micrometer of a microscope and a stopwatch, the ζ potential is calculated by the Smoluchowski equation

$$\zeta = \left(\frac{4\pi\eta}{D}\right)\left(\frac{v}{E/d}\right) = \left(\frac{4\pi\eta}{D}\right)\mu$$

The electrophoretic mobility $u = v/(E/d)$ is the velocity in a potential gradient of 1 v/cm. Particle size and shape do not affect the ζ potential according to the above equation. However, if the particle radius is comparable to δ or smaller (in which case the particles cannot be detected in a microscope), the factor 4 is replaced by 6. The viscosity η and the dielectric constant D refer to the aqueous medium in the double layer and cannot be measured directly.[42] Using the values for water at 25°, expressing the velocity in μm/sec and the electrophoretic mobility in (μm/sec)/(volts/cm), and converting into the appropriate units reduces the Smoluchowski equation to $\zeta = 12.9\,u$, with ζ given in millivolts (mV). If the particle surface has appreciable conductance, the ζ potential calculated by this equation may be low.[25,41,42] Dispersions of hydrophobic particles with ζ potentials below 20–30 mV are frequently unstable and tend to coagulate. On the other hand, values as high as ±180 mV have been reported for the ζ potential.[21,24,41]

The chief experimental precautions in microelectrophoresis measurements are:

1. Electroosmosis causes liquid to flow along the walls of the cell containing the dispersion. This in turn produces a return flow in the center of the cell. The microscope must be focused on the stationary boundary between the two liquid layers flowing in opposite directions in order to measure the true velocity of the particles.
2. Only in very dilute dispersions is it possible to follow the motion of single particles in the microscope field and to measure their velocity. Since the ζ potential depends largely on the nature, ionic strength, and pH of the suspending medium, dispersions should be diluted not with water but with solutions of composition identical to their continuous phase, eg, with their own serum separated by ultrafiltration or centrifugation. The Zeta-Meter is a commercial microelectrophoresis apparatus of easy, fast and reproducible operation.

When the particles cannot be observed individually with a microscope or ultramicroscope, other electrophoresis methods are employed.[24,27,41,43,44] In *moving boundary electrophoresis*, the movement of the boundary formed between a sol or solution and the pure dispersion medium in an electric field is studied. If the disperse phase is colorless, the boundary is located by the refractive index gradient (Tiselius apparatus, used frequently with protein solutions). If several species of particles or solutes with different mobilities are present, each will form a boundary moving with a characteristic velocity. Unlike microelectrophoresis, this method permits the identification of different colloidal components in a mixture, the measurement of the electrophoretic mobility of each, and an estimation of the relative amounts present.

Zone electrophoresis theoretically permits the complete separation of all electrophoretically different components, requires much smaller samples than moving boundary electrophoresis, and can be performed in simpler and less expensive equipment. The method avoids convection by supporting the solution in an inert and porous solid like filter paper, cellulose acetate membrane, agar, starch or polyacrylamide gels cut into strips, or disks or columns of polyacrylamide gel.

A strip of filter paper or gel is saturated with a conducting buffer solution and a few microliters of the solution being analyzed is deposited as a spot or narrow band. A potential difference is applied between the ends of the strip which are

in contact with the electrode compartments. The spot or band spreads and unfolds as each component migrates towards one or the other electrode at a rate determined primarily by its electrophoretic mobility. Evaporation of water due to the heating effect of the electric current may be minimized by immersing the strip in a cooling liquid or sandwiching it between impervious solid sheets. After a sufficient time has elapsed to afford good separation, the strip is removed and dried. The position of the spots or bands corresponding to the individual components is detected by color reactions or radioactive counting.

Zone electrophoresis is applied mainly in analysis and for small-scale preparative separations. It does not permit mobility measurements. Because several samples can be analyzed simultaneously (in parallel strips or gel columns), because only minute amounts of sample are needed, and because the equipment is simple and easy to operate, zone electrophoresis is widely used to study the proteins in blood serum, erythrocytes, lymph and cerebrospinal fluid, saliva, gastric and pancreatic juices and bile.

Immunodiffusion combined with electrophoresis is called *immunoelectrophoresis*.[43,45] The proteins in a fluid, including the antigens, are first separated by gel electrophoresis. A longitudinal trench is then cut along one or both sides of the gel strip near the edge in the direction of the electrophoresis axis. The trench is filled with the antibody solution. On standing, antibody and antigen proteins diffuse in all directions, including toward each other. Precipitation occurs along an elliptical arc (precipitin band) wherever an antigen meets its specific antibody. The precipitin bands are either visible directly or may be developed by staining. Since diseases frequently produce abnormal electrophoretic patterns in body fluids, zone electrophoresis and immunoelectrophoresis are convenient and powerful diagnostic techniques.

Isoelectric focusing[44,46] uses electrophoresis to separate proteins according to their isoelectric points. At pH values equal to their isoelectric points, proteins do not migrate in an electric field because their net charge is zero. In a liquid column on which a pH gradient is imposed, different species arrange themselves so that the protein with the highest isoelectric point will be located nearest to the cathode, which is immersed in the solution of a strong base. The protein with the lowest isoelectric point will be located nearest to the anode, which is immersed in the solution of a strong acid. The other proteins settle into intermediate positions, where the pH values are intermediate and equal to their isoelectric points.

Hydrophilic Dispersions

Most liquid disperse systems of pharmaceutical interest are aqueous. Therefore, most lyophilic colloidal systems discussed below consist of hydrophilic solids dissolved or dispersed in water. Most of the products mentioned below are official in the USP or NF, where more detailed descriptions may be found, also elsewhere in this text.

Hydrophilic colloids can be divided into particulate and soluble materials. The latter are water-soluble linear or branched polymers dissolved molecularly in water. Their aqueous solutions are classified as colloidal dispersions because the individual molecules are in the colloidal particle size range, exceeding 50 or 100 Å. Particulate or corpuscular hydrophilic colloidal dispersions are formed by solids which swell and are peptized in water but whose primary particles do not dissolve or break down into individual molecules or ions. One subdivision of particulate hydrophilic colloids is comprised of dispersions of cross-linked polymers whose linear, uncross-linked analogues are water-soluble.

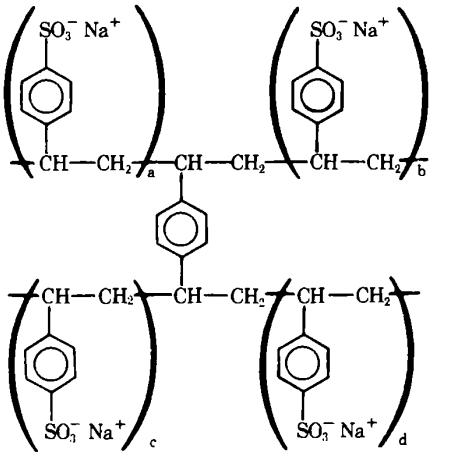
Particulate Hydrophilic Dispersions

The disperse phase of these sols consists of solids which in water swell and break up spontaneously into particles of colloidal dimensions. The disperse particles have high specific surface areas and are, therefore, extensively hydrated. They have characteristic shapes. If the attraction between individual particles is strong, the dispersions have yield values at relatively low solids content.

Bentonite is an aluminum silicate crystallizing in a layer structure (see above), with individual lamellas 9.4 Å thick. Their top and bottom surfaces are sheets of oxygen ions from silica plus an occasional sodium ion neutralizing a silicate ion-exchange site. The clay particles consist of stacks of these lamellas. Water penetrates inside the stacks between lamellas to hydrate the oxygen ions, causing extensive swelling. Bentonite particles in bentonite magma consist of single lamellas and packets of a few lamellas with intercalated water. The specific surface area amounts to several hundred square meters per gram. *Kaolin* also has a layer structure, but does not swell in water because water does not intercalate between individual lattice layers. Kaolin plates dispersed in water are, therefore, much thicker than those of bentonite, *ca* 0.04 to 0.2 μm. In kaolin, hydrated alumina lattice planes alternate with silica planes. Thus, one of the two external surfaces of a kaolin plate consists of a sheet of oxygen ions from silica, the other is a sheet of hydroxide ions from hydrated alumina. Both surfaces are well hydrated. Magnesium aluminum silicate (*Veegum*) is a clay similar to bentonite but contains magnesium; it is white whereas bentonite is gray.

Additional hydrophilic particles producing colloidal dispersions in water are listed below. *Colloidal silicon dioxide* consists of roughly spherical particles covered with siloxane and silanol groups (pages 280–281). *Titanium dioxide* is a white pigment with excellent covering power due to its high refractive index. *Microcrystalline cellulose* (page 279) is hydrophilic because of the hydroxyl and ether groups in the surface of the cellulose crystals. Gelatinous precipitates of hydrophilic compounds such as *aluminum hydroxide gel*, *aluminum phosphate gel*, and *magnesium hydroxide* consist of coarse flocs produced by agglomeration of the colloidal particles formed in the initial stage of the precipitation. They possess large internal surface areas, which is one of the reasons why the first two are used as substrates for adsorbed vaccines and toxoids.

Cross-linked Polymers—The polymers discussed below are polyelectrolytes, ie, they contain ionic groups and would be soluble in water in the absence of cross-linking. For instance, *sodium polystyrene sulfonate* is a copolymer of about 92% styrene and 8% divinylbenzene, which is sulfonated and neutralized to produce the cation-exchange resin



Chains a–b and c–d are water-soluble linear polymer chains. They are cross-linked or bound together via a phenylene group as shown. There are many such cross-links tieing every chain to two or more other chains, so that every atom in a grain of ion-exchange resin is bound to every other atom by primary, covalent bonds. The grains swell in water until the cross-links are strained but do not dissolve, because this would involve the rupture of primary valence bonds. Swelling renders the ion-exchange sites in the interior of a grain accessible to the gastrointestinal fluids. Partial exchange of Na^+ by K^+ followed by excretion of the used resin in the feces reduces hyperkalemia resulting from acute renal failure. Partial replacement of Na^+ by H^+ could reduce acidosis.

Cholestyramine resin is an anion-exchange resin containing the same backbone of cross-linked polystyrene, but substituted with $—CH_2—N^+(CH_3)_3Cl^-$ instead of sodium sulfonate. Part of the chloride anions is exchanged or replaced by bile salt anions, which are thus eliminated in the feces bound to the resin grains rather than reabsorbed. *Colestipol hydrochloride* is another orally administered anion-exchange resin used to increase the fecal excretion of bile salts. It is an extensively cross-linked, insoluble but permeable copolymer made from diethylenetriamine, tetraethylenepentamine, and epichlorohydrin. Strong cation- and anion-exchange resins are used as sustained-release vehicles for basic and acid drugs, respectively (see Chapter 91).

Polycarbophil is a copolymer of acrylic acid cross-linked with a small amount of divinyl glycol. The weakly acidic carboxyl groups are not ionized in the strongly acid environment of the stomach but only in the more nearly neutral intestines. Therefore, swelling by osmotic influx of water occurs mostly in the intestines, where imbibition of water decreases the fluidity of stools associated with diarrhea. Among natural polymers, tragacanth consists of ⅓ of a water-soluble fraction, tragacanthin, and ⅔ of a gel fraction called bassorin which swells in water but does not dissolve. Starch consists of ⅙ of a fraction, soluble in hot water, called amylose. The remainder, amylopectin, merely absorbs water and swells. It owes its insolubility to extensive branching rather than cross-linking.

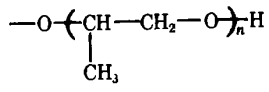
Soluble Polymers as Lyophilic Colloids

Most hydrophilic colloidal systems used in dosage forms are molecular solutions of water soluble, high molecular weight polymers. The polymers are either linear or slightly branched but not cross-linked.

Classifications—According to their origin, water-soluble polymers are divided into three classes. *Natural polymers* include polysaccharides (acacia, agar, heparin sodium, pectin, sodium alginate, tragacanth, xanthan gum) and polypeptides (casein, gelatin, protamine sulfate). Of these, agar and gelatin are only soluble in hot water.

Cellulose derivatives are produced by chemical modification of cellulose obtained from wood pulp or cotton to produce soluble polymers. *Cellulose* is an insoluble, linear polymer of glucose repeat units in the ring or pyranose form joined by β-1,4 glucosidic linkages. Each glucose repeat unit (except for the two terminal ones) contains a primary hydroxyl group on the No 6 carbon and two secondary hydroxyls on No 2 and 3 carbons. The primary hydroxyl is more reactive. Chemical modification of cellulose consists in reactions or substitutions of the hydroxyl groups. The extent of such reactions is expressed as *degree of substitution* (DS), namely, the number of substituted hydroxyl groups per glucose residue. The highest value is DS = 3.0. Fractional values are the rule because the DS is averaged over a multitude of glucose residues. A DS value of 0.6 indicates that some glucose repeat units are unsubstituted while others have one or even two substituents.

Soluble cellulose derivatives are listed below. The DS values correspond to the pharmaceutical grades. The groups shown are the replacements for the hydrogen atoms of the cellulosic hydroxyls. Official derivatives are *methylcellulose* (DS = 1.65–1.93), $—O—CH_3$ and *sodium carboxymethylcellulose* (DS = 0.60–1.00), $—O—CH_2—COO^-Na^+$. *Hydroxyethyl cellulose* (DS ≅ 1.0), $—O\text{-}(CH_2CH_2—O)_n\text{-}H$ and *hydroxypropyl cellulose* (DS ≅ 2.5) are manufactured



by the addition of ethylene oxide and propylene oxide, respectively, to alkali-treated cellulose. The value of n is about 2.0 for the former and not much greater than 1.0 for the latter. *Hydroxypropyl methylcellulose* is prepared by reacting alkali-treated cellulose first with methyl chloride to introduce methoxy groups (DS = 1.1–1.8) and then with propylene oxide to introduce propylene glycol ether groups (DS = 0.1–0.3). In general, the introduction of hydroxypropyl groups into cellulose reduces the water solubility somewhat while promoting the solubility in polar organic solvents like short-chain alcohols, glycols and some ethers.

The molecular weight of native cellulose is so high that soluble derivatives of approximately the same degree of polymerization would dissolve too slowly, and their solutions would be excessively viscous even at concentrations of 1% and less. Controlled degradation is used to break the cellulose chains into shorter segments, reducing the viscosity of the solutions of the corresponding soluble derivatives. Commercial grades of a given cellulose derivative such as sodium carboxymethylcellulose come in various molecular weights or viscosity grades as well as with various degrees of substitution, offering the pharmacist a wide selection.

Official cellulose derivatives which are insoluble in water but soluble in some organic solvents include *ethylcellulose* (DS = 2.2–2.7), $—O—C_2H_5$; *cellulose acetate phthalate* (DS = 1.70 for acetyl and 0.77 for phthalyl); and *pyroxylin* or cellulose nitrate (DS ≅ 2), $—O—NO_2$. *Collodion*, a 4.0% w/v solution of pyroxylin in a mixture of 75% (v/v) ether and 25% (v/v) ethyl alcohol, constitutes a lyophilic colloidal system.

The third class, water soluble *synthetic polymers*, consists mostly of vinyl derivatives including *polyvinyl alcohol*, *povidone* or polyvinylpyrrolidone, and *carbomer* (*Carbopol*), a copolymer of acrylic acid. High molecular weight polyethylene glycols are also called *polyethylene oxides*.

A second classification of hydrophilic polymers is based on their charge. *Nonionic* or uncharged polymers include methylcellulose, hydroxyethyl and hydroxypropyl cellulose, ethylcellulose, pyroxylin, polyethylene oxide, polyvinyl alcohol and povidone. *Anionic* or negatively charged *polyelectrolytes* include the following carboxylated polymers: acacia, alginic acid, pectin, tragacanth, xanthan gum and carbomer at pH values leading to ionization of the carboxyl groups; sodium alginate and sodium carboxymethylcellulose; also polypeptides at pH values above their isoelectric points, eg, sodium caseinate. A stronger acid group is sulfuric acid, which exists as a monoester in agar and heparin and as a monoamide in heparin. *Cationic* or positively charged *polyelectrolytes* are rare. Examples are polypeptides at pH values below their isoelectric points. Protamines are strongly basic due to a high arginine content, with isoelectric points around pH 12, eg protamine sulfate.

Gel Formation—As described in Chapter 20 and illustrated in Fig 20-7*A*, the flexible chains of dissolved polymers interpenetrate and are entangled because of the constant Brownian motion of their segments. The chains writhe and forever change their conformations. Each chain is encased in a sheath of solvent molecules that solvate its functional groups. In the case of aqueous solutions, water molecules

are hydrogen-bonded to the hydroxyl groups of polyvinyl alcohol, hydroxyl groups and ether links of polysaccharides, ether links of polyethylene oxide or polyethylene glycol, amide groups of polypeptides and povidone, and carboxylate groups of anionic polyelectrolytes. The envelope of water of hydration prevents chains segments in close proximity from touching and attracting one another by interchain hydrogen bonds and van der Waals forces as they do in the solid state. The slippage of solvated chains past one another when the solution flows is lubricated by the free solvent between their solvation sheaths.

Factors that lower the hydration of dissolved macromolecules reduce or thin out the sheath of hydration separating adjacent chains. When the hydration is low, contiguous chains tend to attract one another by secondary valence forces including hydrogen bonds and van der Waals forces. Hydrophobic bonding makes an important contribution to interchain attraction between polypeptide chains even in solution. Van der Waals forces and hydrogen bonds thus establish weak and reversible cross-links between chains at their points of contact or entanglement, bringing about phase separation or precipitation.

Most water-soluble polymers have higher solubilities in hot than in cold water and tend to precipitate on cooling, as the sheaths of hydration surrounding adjacent chains become too sparse to prevent interchain attraction. Dilute solutions separate into a solvent phase practically free of polymer and a viscous liquid phase containing practically all of the polymer but still a large excess of solvent. This process is called *simple coacervation* and the polymer-rich liquid phase a *coacervate*.[21,47] If the polymer solution is concentrated enough and/or the temperature low enough, cooling causes the formation of a continuous network of precipitating chains attached to one another through weak cross-links consisting of interchain hydrogen bonds and van der Waals forces at the points of mutual contact. Segments of regularly sequenced polymer chains even associate laterally into crystalline bundles or crystallites. Irregular chain structures as found in random copolymers, randomly substituted cellulose ethers and esters, and highly branched polymers like acacia prevent crystallization during precipitation from solution. Chain entanglements provide the sole temporary cross-links in those cases. The network of associated polymer chains immobilizes the solvent and causes the solution to set to a gel. Gelatinous precipitates or highly swollen flocs may separate when cooling more dilute polymer solutions.

Besides the chemical nature of polymer and solvent, the three most important factors causing phase separation, precipitation and gelation of polymer solutions are temperature, concentration and molecular weight. Lower temperatures, higher concentrations and higher molecular weights promote gelation and produce stronger gels.

For a typical *gelatin*, 10% solutions acquire yield values and begin to gel at about 25°, 20% solutions at about 30° and 30% solutions at about 32°. The *gelation* is reversible: the gels liquefy when heated above these temperatures. Gelation is rarely observed above 34° regardless of concentration, so that gelatin solutions do not gel at 37°. Conversely, gelatin will dissolve readily in water at body temperature. The gelation temperature or gel point of gelatin is highest at the isoelectric point, where the attachment between adjacent chains by coulombic attraction or ionic bonds between carboxylate ions and alkylammonium, guanidinium or imidazolium groups is most extensive. Since the carboxyl groups are not ionized at gastric pH, interchain ionic bonds are practically nonexistent, and interchain attraction is limited to hydrogen bonds and van der Waals forces. The gelation temperature or the melting point of gelatin gels depends more strongly on temperature and concentration than on pH.[48,49] The combination of an acid pH considerably below the isoelectric point and a temperature of 37° completely prevents the gelation of gelatin solutions. Conversely, these two conditions promote rapid dissolution of gelatin capsules in the stomach. Agar and pectic acid solutions set to gels at only a few percent of solids.

Unlike most water-soluble polymers, methylcellulose, hydroxypropyl cellulose and polyethylene oxide are more soluble in cold than in hot water. Their solutions therefore tend to gel on heating (*thermal gelation*).

When dissolving powdered polymers in water, temporary gel formation often slows the process down considerably. As water diffuses into loose clumps of powder, their exterior frequently turns to a cohesive gel of solvated particles encasing dry powder. Such blobs of gel dissolve very slowly because of their high viscosity and the low diffusion coefficient of the macromolecules. Especially for large-scale dissolution, it is helpful to disperse the polymer powder in water before it can agglomerate into lumps of gel. In order to permit dispersion to precede hydration and to prevent temporary gel formation, the polymer powders are dispersed in water at temperatures where the solubility of the polymer is lowest. Most polymer powders, such as sodium carboxymethylcellulose, are dispersed with high shear in *cold* water before the particles can hydrate and swell to sticky gel grains agglomerating into lumps. Once the powder is well dispersed, the solution is heated with moderate shear to about 60° for fastest dissolution. Because methylcellulose hydrates most slowly in hot water, the powder is dispersed with high shear in 1/5 to 1/3 of the required amount of water heated to 80 to 90°. Once the powder is finely dispersed, the rest of the water is added cold or even as ice, and moderate stirring causes prompt dissolution. For maximum clarity, fullest hydration and highest viscosity, the solution should be cooled to 0 to 10° for about an hour.

The following are two alternative methods for preventing the formation of gelatinous lumps upon addition of water. The powder is prewetted with a water-miscible organic solvent such as ethyl alcohol or propylene glycol that does not swell the polymer, in the proportion of from three to five parts solvent to each part of polymer. If other nonpolymeric powdered adjuvants are to be incorporated into the solution, these are dry-blended with the polymer powder. The latter should comprise 1/4 or less of the blend for best results.

A pharmaceutical application of *gelation* in a nonaqueous medium is the manufacture of *Plastibase* or *Jelene* (*Squibb*), which consists of 5% of a low-molecular-weight polyethylene and 95% of mineral oil. The polymer is soluble in mineral oil above 90°, which is close to its melting point. When the solution is cooled below 90°, the polymer precipitates and causes gelation. The mineral oil is immobilized in the network of entangled, and adhering, insoluble polyethylene chains which probably even associate into small crystalline regions. Unlike petrolatum, this gel can be heated to about 60° without substantial loss in consistency.

Large increases in the concentration of polymer solutions may lead to precipitation and gelation. One way of effectively increasing the concentration of aqueous polymer solutions is to add inorganic salts. The salts will bind part of the water of the polymer solution in order to become hydrated. Competition for water of hydration dehydrates the polymer molecules and precipitates them, causing gelation. This phenomenon is called *salting out*. Because of its high solubility in water, ammonium sulfate is often used by biochemists to precipitate and separate proteins from dilute solution. To the pharmacist, salting out usually represents an undesirable problem. It is reversible, however, and subsequent addition of water redissolves the precipitated polymers and liquefies their gels. Salting out may cause the polymer to separate as a concentrated and viscous liquid solution or simple coacervate rather than as a solid gel.

The effectiveness of electrolytes to salt out, precipitate or gel hydrophilic colloidal systems depends on how extensively the electrolytes are hydrated. The *Hofmeister* or *lyotropic series* arranges ions in the order of increasing hydration and increasing effectiveness in salting out hydrophilic colloids. The series, for monovalent cations, is

$$Cs^+ < Rb^+ < NH_4^+ < K^+ < Na^+ < Li^+$$

and for divalent cations,

$$Ba^{2+} < Sr^{2+} < Ca^{2+} < Mg^{2+}$$

This series also arranges the cations in the order of decreasing coagulating power or increasing coagulation values for negative hydrophobic sols (see Table XIV) and of increasing ease of their displacement from cation exchange resins: K^+ displaces Na^+ and Li^+. For anions, the lyotropic series in the order of decreasing coagulating power and decreasing effectiveness in salting out is

$$F^- > \text{citrate}^{3-} > HPO_4^{2-} > \text{tartrate}^{2-} > SO_4^{2-} > \text{acetate}^- > Cl^- > NO_3^- > ClO_3^- > Br^- > ClO_4^- > I^- > CNS^-$$

Iodides and thiocyanates and to a lesser extent bromides and nitrates actually tend to increase the solubility of polymers in water, salting them in.[21,22,24-26] These large polarizable anions destructure water, reducing the extent of hydrogen bonding among water molecules and thereby making more of the hydrogen-bonding capacity of water available to the solute. Most salts except nitrates, bromides, perchlorates, iodides and thiocyanates raise the temperature of precipitation or gelation of most hydrophilic colloidal solutions or their gel melting points. Exceptions among hydrophilic colloids are methylcellulose, hydroxypropyl cellulose and polyethylene oxide whose gelation temperatures or gel points and gel melting points are lowered by salting out.

Hydrophobic aqueous dispersions are coagulated by electrolytes at 0.0001–0.1 *M* concentrations (see Table XIV). Moreover, the coagulation is irreversible, ie, removal of the coagulating salt does not allow the coagulum to be redispersed, because the hydrophobic sols are intrinsically unstable. By contrast, most hydrophilic sols require electrolyte concentrations of 1 *M* or higher for precipitation. Their precipitation or gelation can be reversed, and the polymer redissolved by removing the salt through dialysis or by adding more water. Hydrophilic colloids disperse or dissolve spontaneously in water, and their sols are intrinsically stable.

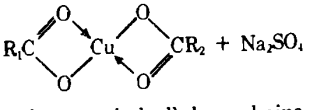
Most of the hydrophilic and water-soluble polymers mentioned above are only slightly soluble or insoluble in alcohol. Addition of alcohol to their aqueous solutions may cause precipitation or gelation because alcohol is a nonsolvent or precipitant, lowering the dielectric constant of the medium, and it tends to dehydrate the hydrophilic solute. Alcohol lowers the concentrations at which electrolytes salt out hydrophilic colloids. Phase separation through the addition of alcohol to an aqueous polymer solution may cause coacervation, ie, the separation of a concentrated viscous liquid phase, rather than precipitation or formation of a gel. Sucrose also competes for water of hydration with hydrophilic colloids, and may cause phase separation. However, most hydrophilic sols tolerate substantially higher concentrations of sucrose than of electrolytes or alcohol. Lower viscosity grades of a given polymer are usually more resistant to electrolytes, alcohol and sucrose than grades of higher viscosity and higher molecular weights.

Whenever hydrophilic colloidal dispersions undergo irreversible precipitation or gelation, chemical reactions are involved. Neither dilution with water nor heating nor attempts to remove the gelling or precipitating agent by washing or dialysis will liquefy those gels or redissolve the gelatinous precipitates formed at lower polymer concentrations. Carboxyl groups are not ionized in strongly acid media. If a polymer owes its solubility to the ionization of these weakly acid groups, reducing the pH of its solution below 3 may lead to precipitation or gelation. This is observed with such carboxylated polymers as many gums, sodium carboxymethylcellulose and carbomer. Hydrogen carboxymethylcellulose swells and disperses but does not dissolve in water. Neutralization to higher pH values returns the carboxyl groups to their ionized state and reverses the gelation or precipitation.

Only the sodium, potassium, ammonium and triethanolammonium salts of carboxylated polymers are well soluble in water. In the case of carboxymethylcellulose, salts with heavy metal cations (silver, copper, mercury, lead) and trivalent cations (aluminum, chromic, ferric) are practically insoluble. Salts with divalent cations, especially of the alkaline earth metals, have borderline solubilities. Generally, higher degrees of substitution tend to increase the tolerance of the carboxymethylcellulose to salts.

Precipitation or gelation occur due to metathesis when inorganic salts of heavy or trivalent cations are mixed with alkali metal salts of carboxylated polymers in solution. For instance, if a soluble copper salt is added to a solution of sodium carboxymethylcellulose, the double decomposition can be written schematically as

$$R_1COO^-Na^+ + R_2COO^-Na^+ + CuSO_4 \longrightarrow$$



$+ Na_2SO_4$

R_1 and R_2 represent two carboxymethylcellulose chains which are cross-linked by a chelated copper ion. Dissociation of the cupric carboxylate complex is negligible.

Particle Phenomena and Coarse Dispersions

The Dispersion Step

The pharmaceutical formulator is concerned primarily with producing a smooth, uniform, easily flowing (pouring or spreading) suspension or emulsion in which dispersion of particles can be effected with minimum expenditure of energy.

In preparing suspensions, particle–particle attractive forces need to be overcome by the high shearing action of such devices as the colloid mill, or by use of surface-active agents. The latter greatly facilitate wetting of lyophobic powders and assist in the removal of surface air that shearing alone may not remove; thus the clumping tendency of the particles is reduced. Moreover, lowering of the surface free energy by the adsorption of these agents directly reduces the thermodynamic driving force opposing dispersion of the particles.

In emulsification shear rates are frequently necessary for dispersion of the internal phase into fine droplets. The shear forces are opposed by forces operating to resist distortion and subsequent breakup of the droplets. Again surface-active agents help greatly by lowering interfacial ten-

sion, which is the primary reversible component resisting droplet distortion. Surface-active agents also may play an important role in determining whether an oil-in-water or a water-in-oil emulsion preferentially survives the shearing action.

Once the process of dispersion begins there develops simultaneously a tendency for the system to revert to an energetically more stable state, manifested by flocculation, coalescence, sedimentation, crystal growth, and caking phenomena. If these physical changes are not inhibited or controlled, successful dispersions will not be achieved or will be lost during shelf life.

Settling and Its Control

In order to control the settling of dispersed material in suspension, the pharmacist must be aware of those physical factors that will affect the rate of sedimentation of particles under ideal and nonideal conditions. He must also be aware of the various coefficients used to express the amount of flocculation in the system and the effect flocculation will have on the structure and volume of the sediment.

Sedimentation Rate

The rate at which particles in a suspension sediment is related to their size and density and the viscosity of the suspension medium. Brownian movement may exert a significant effect, as will the absence or presence of flocculation in the system.

Stokes' Law—The velocity of sedimentation of a uniform collection of spherical particles is governed by Stokes' law, expressed as follows:

$$v = \frac{2r^2(\rho_1 - \rho_2)g}{9\eta} \qquad (35)$$

where v is the terminal velocity in cm/sec, r is the radius of the particles in cm, ρ_1 and ρ_2 are the densities (g/cm^3) of the dispersed phase and the dispersion medium, respectively, g is the acceleration due to gravity (980.7 cm/sec^2) and η is the Newtonian viscosity of the dispersion medium in poises (g/cm sec). Stokes' law holds only if the downward motion of the particles is not sufficiently rapid to cause turbulence. Micelles and small phospholipid vesicles do not settle unless they are subjected to centrifugation.

While conditions in a pharmaceutical suspension are not in strict accord with those laid down for Stokes' law, Eq 35, provides those factors that can be expected to influence the rate of settling. Thus, sedimentation velocity will be reduced by decreasing the particle size, provided the particles are kept in a deflocculated state. The rate of sedimentation will be an inverse function of the viscosity of the dispersion medium. However, too high a viscosity is undesirable, especially if the suspending medium is Newtonian rather than shear-thinning (see Chapter 20), since it then becomes difficult to redisperse material which has settled. It also may be inconvenient to remove a viscous suspension from its container. When the size of particles undergoing sedimentation is reduced to approximately 2 μm, random Brownian movement is observed and the rate of sedimentation departs markedly from the theoretical predictions of Stokes' law. The actual size at which Brownian movement becomes significant depends on the density of the particle as well as the viscosity of the dispersion medium.

Flocculation and Deflocculation—Zeta potential ψ_z is a measurable indication of the potential existing at the surface of a particle. When ψ_z is relatively high (25 mV or more), the repulsive forces between two particles exceed the attractive London forces. Accordingly, the particles are dispersed and are said to be *deflocculated*. Even when brought close together by random motion or agitation, deflocculated particles resist collision due to their high surface potential.

The addition of a preferentially adsorbed ion whose charge is opposite in sign to that on the particle leads to a progressive lowering of ψ_z. At some concentration of the added ion the electrical forces of repulsion are lowered sufficiently that the forces of attraction predominate. Under these conditions the particles may approach each other more closely and form loose aggregates, termed flocs. Such a system is said to be *flocculated*.

Some workers restrict the term *flocculation* to the aggregation brought about by chemical bridging; aggregation involving a reduction of repulsive potential at the double layer is referred to as *coagulation*. Other workers regard flocculation as aggregation in the secondary minimum of the potential energy curve of two interacting particles and coagulation as aggregation in the primary minimum. In the present chapter the term *flocculation* is used for all aggregation processes, irrespective of mechanism.

The continued addition of the flocculating agent can reverse the above process, if the zeta potential increases sufficiently in the opposite direction. Thus, the adsorption of anions onto positively charged deflocculated particles in suspension will lead to flocculation. The addition of more anions can eventually generate a net negative charge on the particles. When this has achieved the required magnitude, deflocculation may occur again. The only difference from the starting system is that the net charge on the particles in their deflocculated state is negative rather than positive.

Table XV—Relative Properties of Flocculated and Deflocculated Particles in Suspension

Deflocculated	Flocculated
1. Particles exist in suspension as separate entities.	Particles form loose aggregates.
2. Rate of sedimentation is slow, since each particle settles separately and particle size is minimal.	Rate of sedimentation is high, since particles settle as a floc, which is a collection of particles.
3. A sediment is formed slowly.	A sediment is formed rapidly.
4. The sediment eventually becomes very closely packed, due to weight of upper layers of sedimenting material. Repulsive forces between particles are overcome and a hard cake is formed which is difficult, if not impossible, to redisperse.	The sediment is loosely packed and possesses a scaffold-like structure. Particles do not bond tightly to each other and a hard, dense cake does not form. The sediment is easy to redisperse, so as to reform the original suspension.
5. The suspension has a pleasing appearance, since the suspended material remains suspended for a relatively long time. The supernatant also remains cloudy, even when settling is apparent.	The suspension is somewhat unsightly, due to rapid sedimentation and the presence of an obvious, clear supernatant region. This can be minimized if the volume of sediment is made large. Ideally, volume of sediment should encompass the volume of the suspension.

Some of the major differences between suspensions of flocculated and deflocculated particles are presented in Table XV.

Effect of Flocculation—In a deflocculated system containing a distribution of particle sizes, the larger particles naturally settle faster than the smaller particles. The very small particles remain suspended for a considerable length of time, with the result that no distinct boundary is formed between the supernatant and the sediment. Even when a sediment becomes discernible, the supernatant remains cloudy.

When the same system is flocculated (in a manner to be discussed later), two effects are immediately apparent. First, the flocs tend to fall together so that a distinct boundary between the sediment and the supernatant is readily observed; second, the supernatant is clear, showing that the very fine particles have been incorporated into the flocs. The initial rate of settling in flocculated systems is determined by the size of the flocs and the porosity of the aggregated mass. Under these circumstances it is perhaps better to use the term *subsidence*, rather than sedimentation.

Quantitative Expressions of Sedimentation and Flocculation

Frequently, the pharmacist needs to assess a formulation in terms of the amount of flocculation in the suspension and to compare this with that found in other formulations. The two parameters commonly used for this purpose are outlined below.

Sedimentation Volume—The *sedimentation volume*, F, is the ratio of the equilibrium volume of the sediment, V_u, to the total volume of the suspension, V_0. Thus,

$$F = V_u/V_0 \tag{36}$$

As the volume of suspension which appears occupied by the sediment increases, the value of F, which normally ranges from nearly 0 to 1, increases. In the system where $F = 0.75$, for example, 75% of the total volume in the container is apparently occupied by the loose, porous flocs forming the sediment. This is illustrated in Fig 19-33. When $F = 1$, no sediment is apparent even though the system is flocculated. This is the ideal suspension for, under these conditions, no sedimentation will occur. Caking also will be absent. Furthermore, the suspension is esthetically pleasing, there being no visible, clear supernatant.

Degree of Flocculation—A better parameter for comparing flocculated systems is the *degree of flocculation*, β, which relates the sedimentation volume of the flocculated suspension, F, to the sedimentation volume of the suspension when deflocculated, F_∞. It is expressed as

$$\beta = F/F_\infty \tag{37}$$

The degree of flocculation is, therefore, an expression of the increased sediment volume resulting from flocculation.

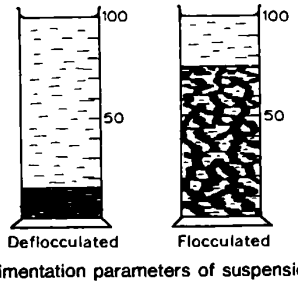


Fig 19-33. Sedimentation parameters of suspensions. Deflocculated suspension: $F_\infty = 0.15$ Flocculated suspension: $F = 0.75$; $\beta = 5.0$.

If, for example, β has a value of 5.0 (Fig 19-33), this means that the volume of sediment in the flocculated system is five times that in the deflocculated state. If a second flocculated formulation results in a value for β of say 6.5, this latter suspension obviously is preferred, if the aim is to produce as flocculated a product as possible. As the degree of flocculation in the system decreases, β approaches unity, the theoretical minimum value.

Suspensions and their Formulation

A pharmaceutical suspension may be defined as a coarse dispersion containing finely divided insoluble material suspended in a liquid medium. Suspension dosage forms are given by the oral route, injected intramusculary or subcutaneously, applied to the skin in topical preparations, and used ophthalmically in the eye. They are an important class of dosage form. Since some products are occasionally prepared in a dry form, to be placed in suspension at the time of dispensing by the addition of an appropriate vehicle, this definition is extended to include these products.

There are certain criteria that a well-formulated suspension should meet. The dispersed particles should be of such a size that they do not settle rapidly in the container. However, in the event that sedimentation occurs, the sediment must not form a hard cake. Rather, it must be capable of redispersion with a minimum effort on the part of the patient. Additionally, the product should be easy to pour, pleasant to take, and resistant to microbial attack.

The three major problem areas associated with suspensions are (1) adequate dispersion of the particles in the vehicle, (2) settling of the dispersed particles, and (3) caking of these particles in the sediment so as to resist redispersion. Much of the following discussion will deal with the factors that influence these processes and the ways in which they can be minimized.

The formulation of a suspension possessing optimal physical stability depends on whether the particles in suspension are to be flocculated or to remain deflocculated. One approach involves use of a structured vehicle to keep deflocculated particles in suspension; a second depends on controlled flocculation as a means of preventing cake formation. A

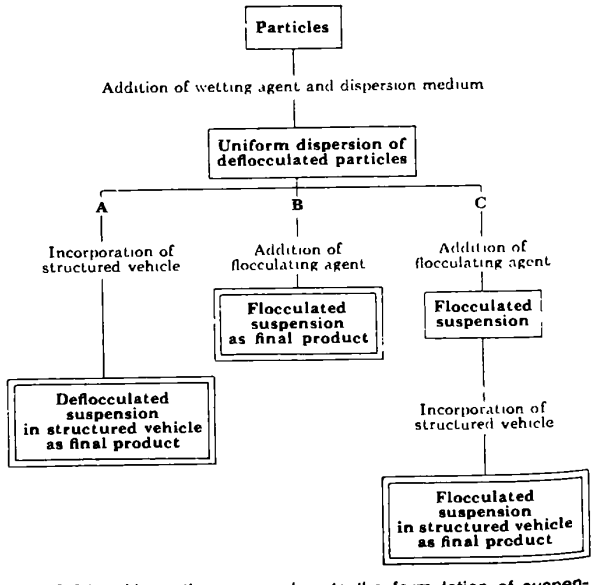


Fig 19-34. Alternative approaches to the formulation of suspensions.

third, a combination of the two previous methods, results in a product with optimum stability. The various schemes are illustrated in Fig 19-34.

Dispersion of Particles—The dispersion step has been discussed earlier in this chapter. Surface-active agents commonly are used as wetting agents; maximum efficiency is obtained when the HLB value lies within the range of 7 to 9. A concentrated solution of the wetting agent in the vehicle may be used to prepare a slurry of the powder; this is diluted with the required amount of vehicle. Alcohol and glycerin may be used sometimes in the initial stages to disperse the particles, thereby allowing the vehicle to penetrate the powder mass.

Only the minimum amount of wetting agent should be used, compatible with producing an adequate dispersion of the particles. Excessive amounts may lead to foaming or impart an undesirable taste or odor to the product. Invariably, as a result of wetting, the dispersed particles in the vehicle are deflocculated.

Structured Vehicles—Structured vehicles are generally aqueous solutions of polymeric materials, such as the hydrocolloids, which are usually negatively charged in aqueous solution. Typical examples are methylcellulose, carboxymethylcellulose, bentonite, and Carbopol. The concentration employed will depend on the consistency desired for the suspension which, in turn, will relate to the size and density of the suspended particles. They function as viscosity-imparting suspending agents and, as such, reduce the rate of sedimentation of dispersed particles.

The rheological properties of suspending agents are considered elsewhere (Chapter 20). Ideally, these form pseudoplastic or plastic systems which undergo shear-thinning. Some degree of thixotropy is also desirable. Non-Newtonian materials of this type are preferred over Newtonian systems because, if the particles eventually settle to the bottom of the container, their redispersion is facilitated by the vehicle thinning when shaken. When the shaking is discontinued, the vehicle regains its original consistency and the redispersed particles are held suspended. This process of redispersion, facilitated by a shear-thinning vehicle, presupposes that the deflocculated particles have not yet formed a cake. If sedimentation and packing have proceeded to the point where considerable caking has occurred, redispersion is virtually impossible.

Controlled Flocculation—When using this approach (see Fig 19-34, B and C), the formulator takes the deflocculated, wetted dispersion of particles and attempts to bring about flocculation by the addition of a flocculating agent; most commonly, these are either electrolytes, polymers, or surfactants. The aim is to *control* flocculation by adding that amount of flocculating agent which results in the maximum sedimentation volume.

Electrolytes are probably the most widely used flocculating agents. They act by reducing the electrical forces of repulsion between particles, thereby allowing the particles to form the loose flocs so characteristic of a flocculated suspension. Since the ability of particles to come together and form a floc depends on their surface charge, zeta potential measurements on the suspension, as an electrolyte is added, provide valuable information as to the extent of flocculation in the system.

This principle is illustrated by reference to the following example, taken from the work of Haines and Martin.[50] Particles of sulfamerazine in water bear a negative charge. The serial addition of a suitable electrolyte, such as aluminum chloride, causes a progressive reduction in the zeta potential of the particles. This is due to the preferential adsorption of the trivalent aluminum cation. Eventually, the zeta potential will reach zero and then become positive as the addition of $AlCl_3$ is continued.

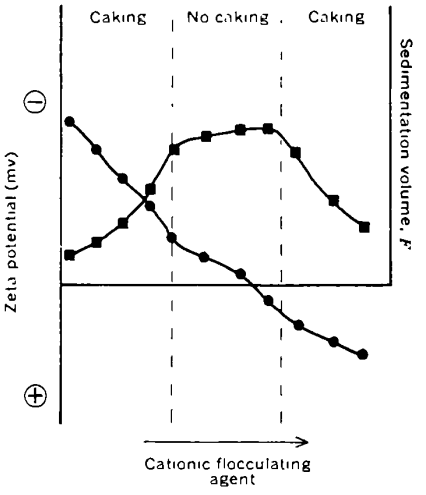


Fig 19-35. Typical relationship between caking, zeta potential and sedimentation volume, as a positively charged flocculating agent is added to a suspension of negatively charged particles. ●: zeta potential; ■: sedimentation volume

If sedimentation studies are run simultaneously on suspensions containing the same range of $AlCl_3$ concentrations, a relationship is observed (Fig 19-35) between the sedimentation volume, F, the presence or absence of caking, and the zeta potential of the particles. In order to obtain a flocculated, noncaking suspension with the maximum sedimentation volume, the zeta potential must be controlled so as to lie within a certain range (generally less than 25 mV). This is achieved by the judicious use of an electrolyte.

A comparable situation is observed when a negative ion such as PO_4^{3-} is added to a suspension of positively charged particles such as bismuth subnitrate. Ionic and nonionic surfactants and lyophilic polymers also have been used to flocculate particles in suspension. Polymers, which act by forming a "bridge" between particles, may be the most efficient additives for inducing flocculation. Thus, it has been shown that the sedimentation volume is higher in suspensions flocculated with an anionic heteropolysaccharide than when electrolytes were used.

Work by Matthews and Rhodes,[51-53] involving both experimental and theoretical studies, has confirmed the formulation principles proposed by Martin and Haines. The suspensions used by Matthews and Rhodes contained 2.5% w/v of griseofulvin as a fine powder together with the anionic surfactant sodium dioxyethylated dodecyl sulfate (10^{-3} molar) as a wetting agent. Increasing concentrations of aluminum chloride were added and the sedimentation height (equivalent to the sedimentation volume, see page 295) and the zeta potential recorded. Flocculation occurred when a concentration of 10^{-3} molar aluminum chloride was reached. At this point the zeta potential had fallen from −46.4 mV to −17.0 mV. Further reduction of the zeta potential, to −4.5 mV by use of 10^{-2} molar aluminum chloride did not increase sedimentation height, in agreement with the principles shown in Fig 19-35.

Matthews and Rhodes then went on to show, by computer analysis, that the DLVO theory (see page 285) predicted the results obtained, namely, that the griseofulvin suspensions under investigation would remain deflocculated when the concentration of aluminum chloride was 10^{-4} molar or less. Only at concentrations in the range of 10^{-3} to 10^{-2} molar aluminum chloride did the theoretical plots show deep primary minima, indicative of flocculation. These occurred at a distance of separation between particles of approximately

50 Å, and led Matthews and Rhodes to conclude that coagulation had taken place in the primary minimum.

Schneider, *et al*[54] have published details of a laboratory investigation (suitable for undergraduates) that combines calculations based on the DLVO theory carried out with an interactive computer program with actual sedimentation experiments performed on simple systems.

Flocculation in Structured Vehicles—The ideal formulation for a suspension would seem to be when flocculated particles are supported in a structured vehicle.

As shown in Fig 19-34 (under C), the process involves dispersion of the particles and their subsequent flocculation. Finally, a lyophilic polymer is added to form the structured vehicle. In developing the formulation, care must be taken to ensure the absence of any incompatibility between the flocculating agent and the polymer used for the structured vehicle. A limitation is that virtually all the structured vehicles in common use are hydrophilic colloids and carry a negative charge. This means that an incompatibility arises if the charge on the particles is originally negative. Flocculation in this instance requires the addition of a positively charged flocculating agent or ion; in the presence of such a material, the negatively charged suspending agent may coagulate and lose its suspendability. This situation does not arise with particles that bear a positive charge, as the negative flocculating agent which the formulator must employ is compatible with the similarly charged suspending agent.

Chemical Stability of Suspensions—Particles that are completely insoluble in a liquid vehicle are unlikely to undergo most chemical reactions leading to degradation. However, most drugs in suspension have a finite solubility, even though this may be of the order of fractions of a microgram per mL. As a result, the material in solution may be susceptible to degradation. However, Tingstad and coworkers[55] developed a simplified method for determining the stability of drugs in suspension. The approach is based on the assumptions that (1) degradation takes place only in the solution and is first order, (2) the effect of temperature on drug solubility and reaction rate conforms with classical theory, and (3) dissolution is not rate-limiting on degradation.

Preparation of Suspensions—The small-scale preparation of suspensions may be readily undertaken by the practicing pharmacist with the minimum of equipment. The initial dispersion of the particles is best carried out by trituration in a mortar, the wetting agent being added in small increments to the powder. Once the particles have been wetted adequately, the slurry may be transferred to the final container. The next step depends on whether the deflocculated particles are to be suspended in a structured vehicle, flocculated, or flocculated and then suspended. Regardless of which of the alternative procedures outlined in Fig 19-34 is employed, the various manipulations can be carried out easily in the bottle, especially if an aqueous solution of the suspending agent has been prepared beforehand.

For a detailed discussion of the methods used in the large-scale production of suspensions, see the relevant section in Chapter 82.

Emulsions in Pharmacy

An emulsion is a dispersed system containing at least two immiscible liquid phases. The majority of conventional emulsions in pharmaceutical use have dispersed particles ranging in diameter from 0.1 to 100 μm. As with suspensions, emulsions are thermodynamically unstable as a result of the excess free energy associated with the surface of the droplets. The dispersed droplets, therefore, strive to come together and reduce the surface area. In addition to this flocculation effect, also observed with suspensions, the dispersed particles can coalesce, or fuse, and this can result in the eventual destruction of the emulsion. In order to minimize this effect a third component, the *emulsifying agent*, is added to the system to improve its stability. The choice of emulsifying agent is critical to the preparation of an emulsion possessing optimum stability. The efficiency of present-day emulsifiers permits the preparation of emulsions which are stable for many months and even years, even though they are thermodynamically unstable.

Emulsions are widely used in pharmacy and medicine, and emulsified materials can possess advantages not observed when formulated in other dosage forms. Thus, certain medicinal agents having an objectionable taste have been made more palatable for oral administration when formulated in an emulsion. The principles of emulsification have been applied extensively in the formulation of dermatological creams and lotions. Intravenous emulsions of contrast media have been developed to assist the physician in undertaking X-ray examinations of the body organs while exposing the patient to the minimum of radiation. Considerable attention has been directed towards the use of sterile, stable intravenous emulsions containing fat, carbohydrate, and vitamins all in one preparation. Such products are administered to patients unable to assimilate these vital materials by the normal oral route.

Emulsions offer potential in the design of systems capable of giving controlled rates of drug release and of affording protection to drugs susceptible to oxidation or hydrolysis. There is still a need for well-characterized dermatological products with reproducible properties, regardless of whether these products are antibacterial, sustained-release, protective, or emollient lotions, creams or ointments. The principle of emulsification is involved in an increasing number of aerosol products.

The pharmacist must be familiar with the types of emulsions and the properties and theories underlying their preparation and stability; such is the purpose of the remainder of this chapter. Microemulsions, which can be regarded as isotropic, swollen micellar systems are discussed in Chapter 83.

Emulsion Type and Means of Detection

A stable emulsion must contain at least three components; namely, the dispersed phase, the dispersion medium, and the emulsifying agent. Invariably, one of the two immiscible liquids is aqueous while the second is an oil. Whether the aqueous or the oil phase becomes the dispersed phase depends primarily on the emulsifying agent used and the relative amounts of the two liquid phases. Hence, an emulsion in which the oil is dispersed as droplets throughout the aqueous phase is termed an oil-in-water, O/W, emulsion. When water is the dispersed phase and an oil the dispersion medium, the emulsion is of the water-in-oil, W/O, type. Most pharmaceutical emulsions designed for oral administration are of the O/W type; emulsified lotions and creams are either O/W or W/O, depending on their use. Butter and salad creams are W/O emulsions.

Recently, so-called *multiple* emulsions have been developed with a view to delaying the release of an active ingredient. In these types of emulsions three phases are present, ie, the emulsion has the form W/O/W or O/W/O. In these

"emulsions within emulsions," any drug present in the innermost phase must now cross two phase boundaries to reach the external, continuous, phase.

It is important for the pharmacist to know the type of emulsion he has prepared or is dealing with, since this can affect its properties and performance. Unfortunately, the several methods available can give incorrect results, and so the type of emulsion determined by one method should always be confirmed by means of a second method.

Dilution Test—This method depends on the fact that an O/W emulsion can be diluted with water and a W/O emulsion with oil. When oil is added to an O/W emulsion or water to a W/O emulsion, the additive is not incorporated into the emulsion and separation is apparent. The test is greatly improved if the addition of the water or oil is observed microscopically.

Conductivity Test—An emulsion in which the continuous phase is aqueous can be expected to possess a much higher conductivity than an emulsion in which the continuous phase is an oil. Accordingly, it frequently happens that when a pair of electrodes, connected to a lamp and an electrical source, are dipped into an O/W emulsion, the lamp lights due to passage of a current between the two electrodes. If the lamp does not light, it is assumed that the system is W/O.

Dye-Solubility Test—The knowledge that a water-soluble dye will dissolve in the aqueous phase of an emulsion while an oil-soluble dye will be taken up by the oil phase provides a third means of determining emulsion type. Thus, if microscopic examination shows that a water-soluble dye has been taken up by the continuous phase, we are dealing with an O/W emulsion. If the dye has not stained the continuous phase, the test is repeated using a small amount of an oil-soluble dye. Coloring of the continuous phase confirms that the emulsion is of the W/O type.

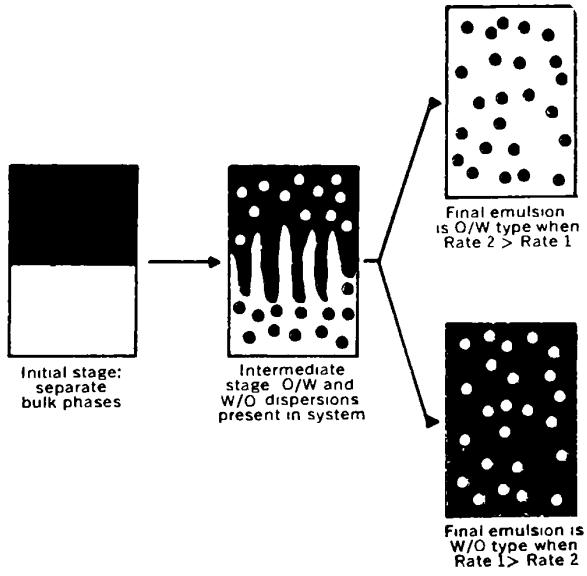


Fig 19-36. Effect of rate of coalescence on emulsion type. Rate 1: O/W coalescence rate; Rate 2: W/O coalescence rate. ●: oil; O: water. For an explanation of Rates 1 and 2, refer to the discussion of Davies on page 304.

Formation and Breakdown of Dispersed Liquid Droplets

An emulsion exists as the result of two competing processes, namely, the dispersion of one liquid throughout another as droplets, and the combination of these droplets to reform the initial bulk liquids. The first process increases the free energy of the system, while the second works to reduce the free energy. Accordingly, the second process is spontaneous and continues until breakdown is complete; ie, the bulk phases are reformed.

It is of little use to form a well-dispersed emulsion if it quickly breaks down. Similarly, unless adequate attention is given to achieving an optimum dispersion during preparation, the stability of an emulsion system may be compromised from the start. Dispersion is brought about by well-designed and well-operated machinery, capable of producing droplets in a relatively short period of time. Such equipment is discussed in Chapter 83. The reversal back to the bulk phases is minimized by utilizing those parameters which influence the stability of the emulsion once it is formed.

Dispersion Process To Form Droplets—Consider two immiscible liquid phases in a test tube. In order to disperse one liquid as droplets within the other, the interface between the two liquids must be disturbed and expanded to a sufficient degree so that "fingers" or threads of one liquid pass into the second liquid, and *vice versa*. These threads are unstable, and become varicosed or beaded. The beads separate and become spherical, as illustrated in Fig 19-36. Depending on the agitation or the shear rate used, larger droplets are also deformed to give small threads, which in turn produce smaller drops.

The time of agitation is important. Thus, the mean size of droplets decreases rapidly in the first few seconds of agitation. The limiting size range is generally reached within 1 to 5 minutes, and results from the number of droplets coalescing being equivalent to the number of new droplets being formed. It is uneconomical to continue agitation any further.

The liquids may be agitated or sheared by several means. Shaking is commonly employed, especially when the components are of low viscosity. Intermittent shaking is frequently more efficient than continual shaking, possibly because the short time interval between shakes allows the thread which is forced across the interface time to break down into drops which are then isolated in the opposite phase. Continuous, rapid agitation tends to hinder this breakdown to form drops. A mortar and pestle is employed frequently in the extemporaneous preparation of emulsions. It is not a very efficient technique and is not used on a large scale. Improved dispersions are achieved by the use of high-speed mixers, blenders, colloid mills and homogenizers. Ultrasonic techniques also have been employed and are described in Chapter 83.

The phenomenon of spontaneous emulsification, as the name implies, occurs without any external agitation. There is, however, an internal agitation arising from certain physicochemical processes that affect the interface between the two bulk liquids. For a description of this process, see Davies and Rideal in the *Bibliography*.

Coalescence of Droplets—Coalescence is a process distinct from flocculation (aggregation), which commonly precedes it. While flocculation is the clumping together of particles, coalescence is the fusing of the agglomerates into a larger drop, or drops. Coalescence is usually rapid when two immiscible liquids are shaken together, since there is no large energy barrier to prevent fusion of drops and reformation of the original bulk phases. When an emulsifying agent is added to the system, flocculation still may occur but coalescence is reduced to an extent depending on the efficacy of the emulsifying agent to form a stable, coherent interfacial film. It is therefore possible to prepare emulsions that are flocculated, yet which do not coalesce. In addition to the interfacial film around the droplets acting as a mechanical

barrier, the drops also are prevented from coalescing by the presence of a thin layer of continuous phase between particles clumped together.

Davies[56] showed the importance of coalescence rates in determining emulsion type; this work is discussed in more detail on page 304.

Emulsifying Agent

The process of coalescence can be reduced to insignificant levels by the addition of a third component—the emulsifying agent or emulsifier. The choice of emulsifying agent is frequently critical in developing a successful emulsion, and the pharmacist should be aware of

The desirable properties of emulsifying agents.
How different emulsifiers act to optimize emulsion stability.
How the type and physical properties of the emulsion can be affected by the emulsifying agent.

Desirable Properties

Some of the desirable properties of an emulsifying agent are that it should

1. Be surface-active and reduce surface tension to below 10 dynes/cm.
2. Be adsorbed quickly around the dispersed drops as a condensed, nonadherent film which will prevent coalescence.
3. Impart to the droplets an adequate electrical potential so that mutual repulsion occurs.
4. Increase the viscosity of the emulsion.
5. Be effective in a reasonably low concentration.

Not all emulsifying agents possess these properties to the same degree; in fact, not every good emulsifier necessarily possesses all these properties. Further, there is no one "ideal" emulsifying agent because the desirable properties of an emulsifier depend, in part, on the properties of the two immiscible phases in the particular system under consideration.

Interfacial Tension—Lowering of interfacial tension is one way in which the increased surface free energy associated with the formation of droplets, and hence surface area, in an emulsion can be reduced (Eq 29). Assuming the droplets to be spherical, it can be shown that

$$\Delta F = \frac{6\gamma V}{d} \qquad (38)$$

where V is the volume of dispersed phase in mL and d is the mean diameter of the particles. In order to disperse 100 mL of oil as 1-μm (10^{-4}-cm) droplets in water when $\gamma_{O/W} = 50$ dynes/cm, requires an energy input of

$$\Delta F = \frac{6 \times 50 \times 100}{1 \times 10^{-4}} = 30 \times 10^7 \text{ ergs}$$

$$= 30 \text{ joules or } 30/4.184 = 7.2 \text{ cal}$$

In the above example the addition of an emulsifier that will reduce γ from 50 to 5 dynes/cm will reduce the surface free energy from 7.2 to around 0.7 cal. Likewise, if the interfacial tension is reduced to 0.5 dyne/cm, a common occurrence, the original surface free energy is reduced a hundredfold. Such a reduction can help to maintain the surface area generated during the dispersion process.

Film Formation—The major requirement of a potential emulsifying agent is that it readily form a film around each droplet of dispersed material. The main purpose of this film—which can be a monolayer, a multilayer, or a collection of small particles adsorbed at the interface—is to form a barrier which prevents the coalescence of droplets that come into contact with one another. For the film to be an efficient barrier, it should possess some degree of surface elasticity and should not thin out and rupture when sandwiched between two droplets. If broken, the film should have the capacity to reform rapidly.

Electrical Potential—The origin of an electrical potential at the surface of a droplet has been discussed earlier in the chapter. Insofar as emulsions are concerned, the presence of a well-developed charge on the droplet surface is significant in promoting stability by causing repulsion between approaching drops. This potential is likely to be greater when an ionized emulsifying agent is employed.

Concentration of Emulsifier—The main objective of an emulsifying agent is to form a condensed film around the droplets of the dispersed phase. An inadequate concentration will do little to prevent coalescence. Increasing the emulsifier concentration above an optimum level achieves little in terms of increased stability. In practice the aim is to use the minimum amount consistent with producing a satisfactory emulsion.

It frequently helps to have some idea of the amount of emulsifier required to form a condensed film, one molecule thick, around each droplet. Suppose we wish to emulsify 50 g of an oil, density = 1.0, in 50 g of water. The desired particle diameter is 1 μm. Thus,

Particle diameter = 1 μm = 1×10^{-4} cm

$$\text{Volume of particle} = \frac{\pi d^3}{6} = 0.524 \times 10^{-12} \text{ cm}^3$$

$$\text{Total number of particles in 50 g} = \frac{50}{0.524 \times 10^{-12}} = 95.5 \times 10^{12}$$

Surface area of each particle = $\pi d^2 = 3.142 \times 10^{-8}$ cm^2

$$\text{Total surface area} = 3.142 \times 10^{-8} \times 95.5 \times 10^{12} = 300 \times 10^4 \text{ cm}^2$$

If the area each molecule occupies at the oil/water interface is 30 Å^2 (30×10^{-16} cm^2), we require

$$\frac{300 \times 10^4}{30 \times 10^{16}} = 1 \times 10^{21} \text{ molecules}$$

A typical emulsifying agent might have a molecular weight of 1000. Thus, the required weight is

$$\frac{1000 \times 10^{21}}{6.023 \times 10^{23}} = 1.66 \text{ g}$$

To emulsify 10 g of oil would require 0.33 g of the emulsifying agent, etc. While the approach is an oversimplification of the problem, it does at least allow the formulator to make a reasonable estimate of the required concentration of emulsifier.

Emulsion Rheology—The emulsifying agent and other components of an emulsion can affect the rheologic behavior of an emulsion in several ways and these are summarized in Table XVI. It should be borne in mind that the droplets of the internal phase are deformable under shear and that the adsorbed layer of emulsifier affects the interactions between adjacent droplets and also between a droplet and the continuous phase.

The means by which the rheological behavior of emulsions can be controlled have been discussed by Rogers.[58]

Mechanism of Action

Emulsifying agents may be classified in accordance with the type of film they form at the interface between the two phases.

Monomolecular Films—Those surface-active agents which are capable of stabilizing an emulsion do so by form-

Table XVI—Factors Influencing Emulsion Viscosity[57]

1. Internal phase
   a. Volume concentration (ϕ); hydrodynamic interaction between globules; flocculation, leading to formation of globule aggregates.
   b. Viscosity (η_i); deformation of globules in shear.
   c. Globule size, and size distribution, technique used to prepare emulsion; interfacial tension between the two liquid phases: globule behavior in shear; interaction with continuous phase; globule interaction.
   d. Chemical constitution.
2. Continuous phase
   a. Viscosity (η_0), and other rheological properties.
   b. Chemical constitution, polarity, pH; potential energy of interaction between globules.
   c. Electrolyte concentration if polar medium.
3. Emulsifying agent
   a. Chemical constitution; potential energy of interaction between globules.
   b. Concentration, and solubility in internal and continuous phases; emulsion type; emulsion inversion; solubilization of liquid phases in micelles.
   c. Thickness of film adsorbed around globules, and its rheological properties, deformation of globules in shear; fluid circulation within globules.
   d. Electroviscous effect.
4. Additional stabilizing agents
   Pigments, hydrocolloids, hydrous oxides; effect on rheologic properties of liquid phases, and interfacial boundary region.

ing a monolayer of adsorbed molecules or ions at the oil/water interface (Fig 19-37). In accordance with Gibbs' law (Eq 29) the presence of an interfacial excess necessitates a reduction in interfacial tension. This results in a more stable emulsion because of a proportional reduction in the surface free energy. Of itself, this reduction is probably not the main factor promoting stability. More significant is the fact that the droplets are surrounded now by a coherent monolayer which prevents coalescence between approaching droplets. If the emulsifier forming the monolayer is ionized, the presence of strongly charged and mutually repelling droplets increases the stability of the system. With unionized, nonionic surface-active agents, the particles may still carry a charge; this arises from adsorption of a specific ion or ions from solution.

Multimolecular Films—Hydrated lyophilic colloids form multimolecular films around droplets of dispersed oil (Fig 19-37). The use of these agents has declined in recent years because of the large number of synthetic surface-active agents available which possess well-marked emulsifying properties. While these hydrophilic colloids are adsorbed at an interface (and can be regarded therefore as "surface-active"), they do not cause an appreciable lowering in surface tension. Rather, their efficiency depends on their ability to form strong, coherent multimolecular films. These act as a coating around the droplets and render them highly resistant to coalescence, even in the absence of a well-developed surface potential. Furthermore, any hydrocolloid not adsorbed at the interface increases the viscosity of the continuous aqueous phase; this enhances emulsion stability.

Solid Particle Films—Small solid particles that are wetted to some degree by both aqueous and nonaqueous liquid phases act as emulsifying agents. If the particles are too hydrophilic, they remain in the aqueous phase; if too hydrophobic, they are dispersed completely in the oil phase. A second requirement is that the particles are small in relation to the droplets of the dispersed phase (Fig 19-37).

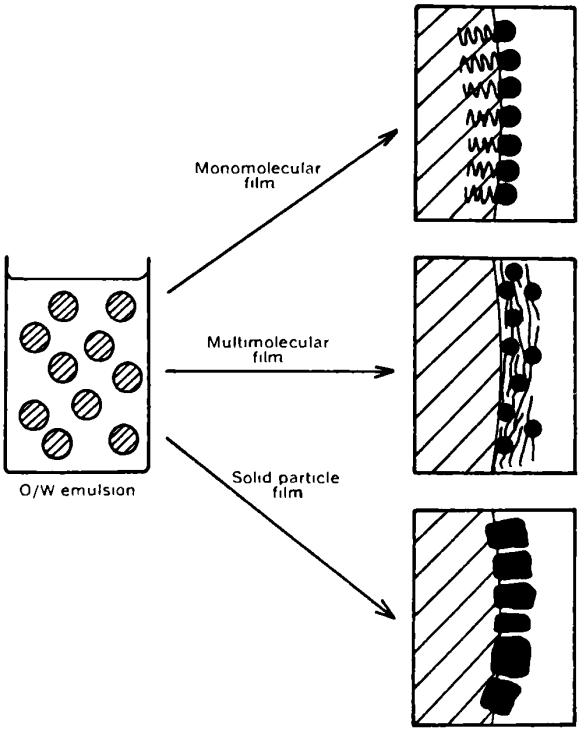


Fig 19-37. Types of films formed by emulsifying agents at the oil/water interface. Orientations are shown for O/W emulsions. ▨: oil; □: water.

Chemical Types

Emulsifying agents may also be classified in terms of their chemical structure; there is some correlation between this classification and that based on the mechanism of action. For example, the majority of emulsifiers forming monomolecular films are synthetic, organic materials. Most of the emulsifiers that form multimolecular films are obtained from natural sources and are organic. A third group is composed of solid particles, invariably inorganic, that form films composed of finely divided solid particles.

Accordingly, the classification adopted divides emulsifying agents into *synthetic*, *natural*, and *finely dispersed solids* (Table XVII). A fourth group, the *auxiliary materials* (Table XVIII), are weak emulsifiers. The agents listed are designed to illustrate the various types available; they are not meant to be exhaustive.

Synthetic Emulsifying Agents—This group of surface-active agents which act as emulsifiers may be subdivided into anionic, cationic, and nonionic, depending on the charge possessed by the surfactant.

Anionics—In this subgroup the surfactant ion bears a negative charge. The potassium, sodium, and ammonium salts of lauric and oleic acid are soluble in water and are good O/W emulsifying agents. They do, however, have a disagreeable taste and are irritating to the gastrointestinal tract; this limits them to emulsions prepared for external use. Potassium laurate, a typical example, has the structure

$$CH_3(CH_2)_{10}COO^- \quad K^+$$

Solutions of alkali soaps have a high pH; they start to precipitate out of solution below pH 10 because the unionized fatty acid is now formed, and this has a low aqueous solubility. Further, the free fatty acid is ineffective as an emulsifier and so emulsions formed from alkali soaps are not stable at pH values less than about 10.

The calcium, magnesium and aluminum salts of fatty acids, often termed the metallic soaps, are water insoluble and result in W/O emulsions.

Table XVII—Classification of Emulsifying Agents

Type	Type of film	Examples
Synthetic (surface-active agents)	Monomolecular	*Anionic* Soaps Potassium laurate Triethanolamine stearate Sulfates Sodium lauryl sulfate Alkyl polyoxyethylene sulfates Sulfonates Dioctyl sodium sulfosuccinate *Cationic:* Quaternary ammonium compounds Cetyltrimethylammonium bromide Lauryldimethylbenzylammonium chloride *Nonionic.* Polyoxyethylene fatty alcohol ethers Sorbitan fatty acid esters Polyoxyethylene sorbitan fatty acid esters
Natural	Multimolecular	*Hydrophilic colloids.* Acacia Gelatin
	Monomolecular	Lecithin Cholesterol
Finely divided solids	Solid particle	*Colloidal clays* Bentonite Veegum *Metallic hydroxides:* Magnesium hydroxide

Table XVIII—Auxiliary Emulsifying Agents[55]

Product	Source and composition	Principal use
Bentonite	Colloidal hydrated aluminum silicate	Hydrophilic thickening agent and stabilizer for O/W and W/O lotions and creams
Cetyl alcohol	Chiefly $C_{16}H_{33}OH$	Lipophilic thickening agent and stabilizer for O/W lotions and ointments
Glyceryl monostearate	$C_{17}H_{35}COOCH_2CHOHCH_2OH$	Lipophilic thickening agent and stabilizer for O/W lotions and ointments
Methylcellulose	Series of methyl esters of cellulose	Hydrophilic thickening agent and stabilizer for O/W emulsions; weak O/W emulsifier
Sodium alginate	The sodium salt of alginic acid, a purified carbohydrate extracted from giant kelp	Hydrophilic thickening agent and stabilizer for O/W emulsions
Sodium carboxymethylcellulose	Sodium salt of the carboxymethyl esters of cellulose	Hydrophilic thickening agent and stabilizer for O/W emulsions
Stearic acid	A mixture of solid acids from fats, chiefly stearic and palmitic	Lipophilic thickening agent and stabilizer for O/W lotions and ointments. Forms a true emulsifier when reacted with an alkali
Stearyl alcohol	Chiefly $C_{18}H_{37}OH$	Lipophilic thickening agent and stabilizer for O/W lotions and ointments
Veegum	Colloidal magnesium aluminum silicate	Hydrophilic thickening agent and stabilizer for O/W lotions and creams

Another class of soaps are salts formed from a fatty acid and an organic amine such as triethanolamine. While these O/W emulsifiers are also limited to external preparations, their alkalinity is considerably less than that of the alkali soaps and they are active as emulsifiers down to around pH 8. These agents are less irritating than the alkali soaps.

Sulfated alcohols are neutralized sulfuric acid esters of such fatty alcohols as lauryl and cetyl alcohol. These compounds are an important group of pharmaceutical surfactants. They are used chiefly as wetting agents, although they do have some value as emulsifiers, particularly, when used in conjunction with an auxiliary agent. A frequently used compound is sodium lauryl sulfate.

$$CH_3(CH_2)_{10}CH_2OSO_3^-\ \ Na^+$$

Sulfonates are a class of compounds in which the sulfur atom is connected directly to the carbon atom, giving the general formula

$$CH_3(CH_2)_nCH_2SO_3^-\ \ Na^+$$

Sulfonates have a higher tolerance to calcium ions and do not hydrolyze as readily as the sulfates. A widely used surfactant of this type is dioctyl sodium sulfosuccinate.

Cationics—The surface activity in this group resides in the positively charged cation. These compounds have marked bactericidal properties. This makes them desirable in emulsified anti-infective products such as skin lotions and creams. The pH of an emulsion prepared with a cationic emulsifier lies in the pH 4–6 range. Since this includes the normal pH of the skin, cationic emulsifiers are advantageous in this regard also.

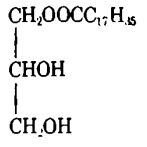
Cationic agents are weak emulsifiers and are generally formulated with a stabilizing or auxiliary emulsifying agent such as cetostearyl alcohol. The only group of cationic agents used extensively as emulsifying agents are the quaternary ammonium compounds. An example is cetyltrimethylammonium bromide.

$$CH_3(CH_2)_{14}CH_2N^+(CH_3)_3\ Br^-$$

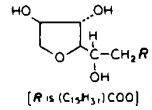
Cationic emulsifiers should not be used in the same formulation with anionic emulsifiers as they will interact. While the incompatibility may not be immediately apparent as a precipitate, virtually all of the desired antibacterial activity will generally have been lost.

Nonionics—These undissociated surfactants find widespread use as emulsifying agents when they possess the proper balance of hydrophilic and lipophilic groups within the molecule. Their popularity is based on the fact that, unlike the anionic and cationic types, nonionic emulsifiers are not susceptible to pH changes and the presence of electrolytes. The number of nonionic agents available is legion; the most frequently used are the glyceryl esters, polyoxyethylene glycol esters and ethers, and the sorbitan fatty acid esters and their polyoxyethylene derivatives.

A glyceryl ester, such as glyceryl monostearate, is too lipophilic to serve as a good emulsifier; it is widely used as an auxiliary agent (Table XVIII) and has the structure

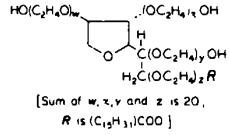


Sorbitan fatty acid esters, such as sorbitan monopalmitate



[R is $(C_{15}H_{31})COO$]

are nonionic oil-soluble emulsifiers that promote W/O emulsions. The polyoxyethylene sorbitan fatty acid esters, such as polyoxyethylene sorbitan monopalmitate, are hydrophilic water-soluble derivatives that favor O/W emulsions.



[Sum of w, x, y and z is 20, R is $(C_{15}H_{31})COO$]

Polyoxyethylene glycol esters, such as the monostearate, $C_{17}H_{35}COO(CH_2OCH_2)_nH$, also are used widely.

Very frequently, the best results are obtained from blends of nonionic emulsifiers. Thus, an O/W emulsifier customarily will be used in an emulsion with a W/O emulsifier. When blended properly, the nonionics produce fine-textured stable emulsions.

Natural Emulsifying Agents—Of the numerous emulsifying agents derived from natural (ie, plant and animal) sources, consideration will be given only to acacia, gelatin, lecithin, and cholesterol. Many other natural materials are only sufficiently active to function as auxiliary emulsifying agents or stabilizers.

Acacia is a carbohydrate gum that is soluble in water and forms O/W emulsions. Emulsions prepared with acacia are stable over a wide pH range. Because it is a carbohydrate it is necessary to preserve acacia emulsions against microbial attack by the use of a suitable preservative. The gum can be precipitated from aqueous solution by the addition of high concentrations of electrolytes or solvents less polar than water, such as alcohol.

Gelatin, a protein, has been used for many years as an emulsifying agent. Gelatin can have two isoelectric points, depending on the method of preparation. So-called Type A gelatin, derived from an acid-treated precursor, has an isoelectric point of between pH 7 and 9. Type B gelatin, obtained from an alkali-treated precursor, has an isoelectric point of approximately pH 5. Type A gelatin acts best as an emulsifier around pH 3, where it is positively charged; on the other hand, Type B gelatin is best used around pH 8, where it is negatively charged. The question as to whether the gelatin is positively or negatively charged is fundamental to the stability of the emulsion when other charged emulsifying agents are present. In order to avoid an incompatibility, all emulsifying agents should carry the same sign. Thus, if gums (such as tragacanth, acacia or agar) which are negatively charged are to be used with gelatin, Type B material should be used at an alkaline pH. Under these conditions the gelatin is similarly negatively charged.

Lecithin is a phospholipid which, because of its strongly hydrophilic nature, produces O/W emulsions. It is liable to microbial attack and tends to darken on storage.

Cholesterol is a major constituent of wool alcohols, obtained by the saponification and fractionation of wool fat. It is cholesterol that gives wool fat its capacity to absorb water and form a W/O emulsion.

Finely Dispersed Solids—This group of emulsifiers forms particulate films around the dispersed droplets and produces emulsions which, while coarse-grained, have considerable physical stability. It appears possible that any solid can act as an emulsifying agent of this type, provided it is reduced to a sufficiently fine powder. In practice the group of compounds used most frequently are the colloidal clays.

Several colloidal clays find application in pharmaceutical emulsions; the most frequently used are bentonite, a colloidal aluminum silicate, and Veegum (*Vanderbilt*), a colloidal magnesium aluminum silicate.

Bentonite is a white to gray, odorless, and tasteless powder that swells in the presence of water to form a translucent suspension with a pH of about 9. Depending on the sequence of mixing it is possible to prepare both O/W and W/O emulsions. When an O/W emulsion is desired, the bentonite is first dispersed in water and allowed to hydrate so as to form a magma. The oil phase is then added gradually with constant trituration. Since the aqueous phase is always in excess, the O/W emulsion type is favored. To prepare a W/O emulsion, the bentonite is first dispersed in oil; the water is then added gradually.

While Veegum is used as a solid particle emulsifying agent, it is employed most extensively as a stabilizer in cosmetic lotions and creams. Concentrations of less than 1% Veegum will stabilize an emulsion containing anionic or nonionic emulsifying agents.

Auxiliary Emulsifying Agents—Included under this heading are those compounds which are normally incapable themselves of forming stable emulsions. Their main value lies in their ability to function as thickening agents and thereby help stabilize the emulsion. Agents in common use are listed in Table XVIII.

Emulsifying Agents and Emulsion Type

For a molecule, ion, colloid, or particle to be active as an emulsifying agent, it must have some affinity for the interface between the dispersed phase and the dispersion medium. With the mono- and multilayer films the emulsifier is in solution and, therefore, must be soluble to some extent in one or both of the phases. At the same time it must not be overly soluble in either phase, otherwise it will remain in the bulk of that phase and not be adsorbed at the interface. This balanced affinity for the two phases also must be evident with finely divided solid particles used as emulsifying agents. If their affinity, as evidenced by the degree to which they are wetted, is either predominantly hydrophilic or hydrophobic, they will not function as effective wetting agents.

The great majority of the work on the relation between

Table XIX—Approximate HLB Values for a Number of Emulsifying Agents

Generic or chemical name	HLB
Sorbitan trioleate	1.8
Sorbitan tristearate	2.1
Propylene glycol monostearate	3.4
Sorbitan sesquioleate	3.7
Glycerol monostearate (non self-emulsifying)	3.8
Sorbitan monooleate	4.3
Propylene glycol monolaurate	4.5
Sorbitan monostearate	4.7
Glyceryl monostearate (self-emulsifying)	5.5
Sorbitan monopalmitate	6.7
Sorbitan monolaurate	8.6
Polyoxyethylene-4-lauryl ether	9.5
Polyethylene glycol 400 monostearate	11.6
Polyoxyethylene-4-sorbitan monolaurate	13.3
Polyoxyethylene-20-sorbitan monooleate	15.0
Polyoxyethylene-20-sorbitan monopalmitate	15.6
Polyoxyethylene-20-sorbitan monolaurate	16.7
Polyoxyethylene-40-stearate	16.9
Sodium oleate	18.0
Sodium lauryl sulfate	40.0

emulsifier and emulsion type has been concerned with surface-active agents that form interfacial monolayers. The present discussion, therefore, will concentrate on this class of agents.

Hydrophile–Lipophile Balance—As the emulsifier becomes more hydrophilic, its solubility in water increases and the formation of an O/W emulsion is favored. Conversely, W/O emulsions are favored with the more lipophilic emulsifiers. This led to the concept that the type of emulsion is related to the balance between hydrophilic and lipophilic solution tendencies of the surface-active emulsifying agent.

Griffin[59] developed a scale based on the balance between these two opposing tendencies. This so-called *HLB scale* is a numerical scale, extending from 1 to approximately 50. The more hydrophilic surfactants have high HLB numbers (in excess of 10), while surfactants with HLB numbers from 1 to 10 are considered to be lipophilic. Surfactants with a proper balance in their hydrophilic and lipophilic affinities are effective emulsifying agents since they concentrate at the oil/water interface. The relationship between HLB values and the application of the surface-active agent is shown in Table XV. Some commonly used emulsifiers and their HLB numbers are listed in Table XIX. The utility of the HLB system in rationalizing the choice of emulsifying agents when formulating an emulsion will be discussed in a later section.

Rate of Coalescence and Emulsion Type—Davies[56] indicated that the type of emulsion produced in systems prepared by shaking is controlled by the relative coalescence rates of oil droplets dispersed in the oil. Thus, when a mixture of oil and water is shaken together with an emulsifying agent, a multiple dispersion is produced initially which contains oil dispersed in water and water dispersed in oil (Fig 19-36). The type of the final emulsion which results depends on whether the water or the oil droplets coalesce more rapidly. If the O/W coalescence rate (Rate 1) is much greater than W/O coalescence rate (Rate 2), a W/O emulsion is formed since the dispersed water droplets are more stable than the dispersed oil droplets. Conversely, if Rate 2 is significantly faster than Rate 1, the final emulsion is an O/W dispersion because the oil droplets are more stable.

According to Davies, the rate at which oil globules coalesce when dispersed in water is given by the expression

$$\text{Rate 1} = C_1 e^{-W_1/RT} \qquad (39)$$

The term C_1 is a collision factor which is directly proportional to the phase volume of the oil relative to the water, and is an inverse function of the viscosity of the continuous phase (water). W_1 defines an energy barrier made up of several contributing factors that must be overcome before coalescence can take place. First, it depends on the electrical potential of the dispersed oil droplets, since this affects repulsion. Second, with an O/W emulsion, the hydrated layer surrounding the polar portion of emulsifying agent must be broken down before coalescence can occur. This hydrated layer is probably around 10 Å thick with a consistency of butter. Finally, the total energy barrier depends on the fraction of the interface covered by the emulsifying agent.

Equation 40 describes the rate of coalescence of water globules dispersed in oil, namely

$$\text{Rate 2} = C_2 e^{-W_2/RT} \qquad (40)$$

Here, the collision factor C_2 is a function of the water/oil phase volume ratio divided by the viscosity of the oil phase. The energy barrier W_2 is, as before, related to the fraction of the interface covered by the surface-active agent. Another contributing factor is the number of $—CH_2—$ groups in the emulsifying agent; the longer the alkyl chain of the emulsifier, the greater the gap that has to be bridged if one water droplet is to combine with a second drop.

Davies[56] showed that the HLB concept is related to the distribution characteristics of the emulsifying agent between the two immiscible phases. An emulsifier with an HLB of less than 7 will be preferentially soluble in the oil phase and will favor formation of a W/O emulsion. Surfactants with an HLB value in excess of 7 will be distributed in favor of the aqueous phase and will promote O/W emulsions.

Preparation of Emulsions

Several factors must be taken into account in the successful preparation and formulation of emulsified products. Usually, the type of emulsion (ie, O/W or W/O) is specified; if not, it probably will be implied from the anticipated use of the product. The formulator's attention is focused primarily on the selection of the emulsifying agent, or agents, necessary to achieve a satisfactory product. No incompatibilities should occur between the various emulsifiers and the several components commonly present in pharmaceutical emulsions. Finally, the product should be prepared in such a way as not to prejudice the formulation.

Selection of Emulsifying Agents

The selection of the emulsifying agent, or agents, is of prime importance in the successful formulation of an emulsion. In addition to its emulsifying properties, the pharmacist must ensure that the material chosen is nontoxic and that the taste, odor, and chemical stability are compatible with the product. Thus, an emulsifying agent which is entirely suitable for inclusion in a skin cream may be unacceptable in the formulation of an oral preparation due to its potential toxicity. This consideration is most important when formulating intravenous emulsions.

The HLB System—With the increasing number of available emulsifiers, particularly the nonionics, the selection of emulsifiers for a product was essentially a trial-and-error procedure. Fortunately, the work of Griffin[59,60] provided a logical means of selecting emulsifying agents. Griffin's method, based on the balance between the hydrophilic and lipophilic portions of the emulsifying agent, is now widely used and has come to be known as the *HLB system*. It is used most in the rational selection of combinations of non-

Table XX—Relationship between HLB Range and Surfactant Application

HLB range	Use
0–3	Antifoaming agents
4–6	W/O emulsifying agents
7–9	Wetting agents
8–18	O/W emulsifying agents
13–15	Detergents
10–18	Solubilizing agents

Table XXI—Required HLB Values for Some Common Emulsion Ingredients

Substance	W/O	O/W
Acid, stearic	...	17
Alcohol, cetyl	...	13
Lanolin, anhydrous	8	15
Oil, cottonseed	...	7.5
mineral oil, light	4	10–12
mineral oil, heavy	4	10.5
Wax, beeswax	5	10–16
microcrystalline	...	9.5
paraffin	...	9

ionic emulsifiers, and we shall limit our discussion accordingly.

As shown in Table XX, if an O/W emulsion is required, the formulator should use emulsifiers with an HLB in the range of 8–18. Emulsifiers with HLB values in the range of 4–6 are given consideration when a W/O emulsion is desired. Some typical examples are given in Table XIX.

Another factor is the presence or absence of any polarity in the material being emulsified, since this will affect the polarity required in the emulsifier. Again, as a result of extensive experimentation, Griffin evolved a series of "required HLB" values; ie, the HLB value required by a particular material if it is to be emulsified effectively. Some values for oils and related materials are contained in Table XXI. Naturally, the required HLB value differs depending on whether the final emulsion is O/W or W/O.

Fundamental to the utility of the HLB concept is the fact that the HLB values are algebraically additive. Thus, by using a low HLB surfactant with one having a high HLB it is possible to prepare blends having HLB values intermediate between those of the two individual emulsifiers. Naturally, one should not use emulsifiers that are incompatible. The following formula should serve as an example.

O/W Emulsion

Liquid petrolatum (Required HLB 10.5)	50 g
Emulsifying agents	5 g
Sorbitan monooleate (HLB 4.3)	
Polyoxyethylene 20 sorbitan monoleate (HLB 15.0)	
Water, qs	100 g

By simple algebra it can be shown that 4.5 parts by weight of sorbitan monooleate blended with 6.2 parts by weight of polyoxyethylene 20 sorbitan monooleate will result in a mixed emulsifying agent having the required HLB of 10.5. Since the formula calls for 5 g, the required weights are 2.1 g and 2.9 g, respectively. The oil-soluble sorbitan monooleate is dissolved in the oil and heated to 75°; the water-soluble polyoxyethylene 20 sorbitan monooleate is added to the aqueous phase which is heated to 70°. At this point the oil phase is mixed with the aqueous phase and the whole stirred continuously until cool.

The formulator is not restricted to these two agents to produce a blend with an HLB of 10.5. Table XXII shows

Table XXII—Nonionic Blends having HLB Values of 10.5

Surfactant blend	HLB	Required amounts (%) to give HLB = 10 5
Sorbitan tristearate	2.1	34.4
Polyoxethylene 20 sorbitan monostearate	14.9	65.6
Sorbitan monopalmitate	6.7	57.3
Polyoxyethylene 20 sorbitan monopalmitate	15.6	42.7
Sorbitan sesquioleate	3.7	48.5
Polyoxyethylene lauryl ether	16.9	51.5

the various proportions required, using other pairs of emulsifying agents, to form a blend of HLB 10.5. When carrying out preliminary investigations with a particular material to be emulsified, it is advisable to try several pairs of emulsifying agents. Based on an evaluation of the emulsions produced, it becomes possible to choose the best combination.

Occasionally, the required HLB of the oil may not be known, in which case it becomes necessary to determine this parameter. Various blends are prepared to give a wide range of HLB mixtures and emulsions are prepared in a standardized manner. The HLB of the blend used to emulsify the best product, selected on the basis of physical stability, is taken to be the required HLB of the oil. The experiment should be repeated using another combination of emulsifiers to confirm the value of the required HLB of the oil to within, say, ±1 HLB unit.

There are methods for finding the HLB value of a new surface-active agent. Griffin[60] developed simple equations which can be used to obtain an estimate with certain compounds. It has been shown that the ability of a compound to spread at a surface is related to its HLB. In another approach a linear relation between HLB and the logarithm of the dielectric constant for a number of nonionic surfactants has been observed. An interesting approach has been developed by Davies[56] and is related to his studies on the relative rates of coalescence of O/W and W/O emulsions (page 304). According to Davies, hydrophilic groups on the surfactant molecule make a positive contribution to the HLB number, whereas lipophilic groups exert a negative effect. Davies calculated these contributions and termed them HLB Group Numbers (Table XXIII). Provided the molecular structure of the surfactant is known, one simply adds the various group numbers in accordance with the following formula:

Table XXIII—HLB Group Numbers[61]

	Group number
Hydrophilic groups	
$-SO_4^-Na^+$	38.7
$-COO^-K^+$	21.1
$-COO^-Na^+$	19.1
N (tertiary amine)	9.4
Ester (sorbitan ring)	6.8
Ester (free)	2.4
—COOH	2.1
Hydroxyl (free)	1.9
—O—	1.3
Hydroxyl (sorbitan ring)	0.5
Lipophilic groups	
—CH—	
$-CH_2-$	
CH_3-	−0.475
=CH—	
Derived groups	
$-(CH_2-CH_2-O)-$	+0.33
$-(CH_2-CH_2-CH_2-O)-$	−0.15

$$HLB = \Sigma(\text{hydrophilic group numbers}) - m(\text{group number}/\text{—}CH_2\text{— group}) + 7$$

where m is the number of —CH_2— groups present in the surfactant. Poor agreement is found between the HLB values calculated by the use of group numbers and the HLB values obtained using the simple equations developed by Griffin. However, the student should realise that the absolute HLB values *per se* are of limited significance. The utility of the HLB approach (using values calculated by either Griffin's or Davies' equations) is to (i) provide the formulator with an idea of the relative balance of hydrophilicity and lipophilicity in a particular surfactant, and (ii) relate that surfactant's emulsifying and solubilizing properties to other surfactants. The formulator still needs to confirm experimentally that a particular formulation will produce a stable emulsion.

Later, Davies and Rideal[61] attempted to relate HLB to the C_{water}/C_{oil} partition coefficient and found good agreement for a series of sorbitan surfactants. Schott[62] showed, however, that the method does not apply to polyoxyethylated octylphenol surfactants. Schott concluded that "so far, the search for a universal correlation between HLB and another property of the surfactant which could be determined more readily than HLB has not been successful."

The HLB system gives no information as to the *amount* of emulsifier required. Having once determined the correct blend, the formulator must prepare another series of emulsions, all at the same HLB, but containing increasing concentrations of the emulsifier blend. Usually, the minimum concentration giving the desired degree of physical stability is chosen.

Mixed Emulsifying Agents—Emulsifying agents are frequently used in combination since a better emulsion usually is obtained. This enhancement may be due to several reasons, one or more of which may be operative in any one system. Thus, the use of a blend or mixture of emulsifiers may (1) produce the required hydrophile–lipophile balance in the emulsifier, (2) enhance the stability and cohesiveness of the interfacial film, and (3) affect the consistency and feel of the product.

The first point has been considered in detail in the previous discussion of the HLB system.

With regard to the second point, Schulman and Cockbain in 1940 showed that combinations of certain amphiphiles formed stable films at the air/water interface. It was postulated that the complex formed by these two materials (one, oil-soluble; the other, water-soluble) at the air/water interface was also present at the O/W interface. This interfacial complex was held to be responsible for the improved stability. For example, sodium cetyl sulfate, a moderately good O/W emulsifier, and elaidyl alcohol or cholesterol, both stabilizers for W/O emulsions, show evidence of an interaction at the air/water interface. Furthermore, an O/W emulsion prepared with sodium cetyl sulfate and elaidyl alcohol is much more stable than an emulsion prepared with sodium cetyl sulfate alone.

Elaidyl alcohol is the *trans* isomer. When oleyl alcohol, the *cis* isomer, is used with sodium cetyl sulfate, there is no evidence of complex formation at the air/water interface. Significantly, this combination does not produce a stable O/W emulsion either. Such a finding strongly suggests that a high degree of molecular alignment is necessary at the O/W interface to form a stable emulsion.

Finally, some materials are added primarily to increase the consistency of the emulsion. This may be done to increase stability or improve emolliency and feel. Examples include cetyl alcohol, stearic acid and beeswax.

When using combinations of emulsifiers, care must be taken to ensure their compatibility, as charged emulsifying agents of opposite sign are likely to interact and coagulate when mixed.

Small-Scale Preparation

Mortar and Pestle—This approach invariably is used only for those emulsions that are stabilized by the presence of a multimolecular film (eg, acacia, tragacanth, agar, chondrus) at the interface. There are two basic methods for preparing emulsions with the mortar and pestle. These are the *Wet Gum* (or so-called *English*) *Method* and the *Dry Gum* (or so-called *Continental*) *Method*.

The Wet Gum Method—In this method the emulsifying agent is placed in the mortar and dispersed in water to form a mucilage. The oil is added in small amounts with continuous trituration, each portion of the oil being emulsified before adding the next increment. Acacia is the most frequently used emulsifying agent when preparing emulsions with the mortar and pestle. When emulsifying a fixed oil, the optimum ratio of oil : water : acacia to prepare the initial emulsion is 4 : 2 : 1. Thus, the preparation of 60 mL of a 40% cod liver oil emulsion requires the following:

Cod liver oil	24 g
Acacia	6 g
Water, qs	60 mL

The acacia mucilage is formed by adding 12 mL of water to the 6 g of acacia in the mortar and triturating. The 24 g of oil is added in increments of 1–2 g and dispersed. The product at this stage is known as the *primary emulsion*, or *nucleus*. The primary emulsion should be triturated for at least 5 min, after which sufficient water is added to produce a final volume of 60 mL.

The Dry Gum Method—In this method, preferred by most pharmacists, the gum is added to the oil, rather than the water as with the wet gum method. Again, the approach is to prepare a primary emulsion from which the final product can be obtained by dilution with the continuous phase. If the emulsifier is acacia and a fixed oil is to be emulsified, the ratio of oil : water : gum is again 4 : 2 : 1.

Provided dispersion of the acacia in the oil is adequate, the dry gum method can almost be guaranteed to produce an acceptable emulsion. Because there is no incremental addition of one of the components, the preparation of an emulsion by this method is rapid.

With both methods the oil : water : gum ratio may vary, depending on the type of oil to be emulsified and the emulsifying agent used. The usual ratios for tragacanth and acacia are shown in Table XXIV.

The preparation of emulsions by both the wet and dry gum methods can be carried out in a bottle rather than a mortar and pestle.

Other Methods—An increasing number of emulsions are being formulated with synthetic emulsifying agents, especially of the nonionic type. The components in such a for-

Table XXIV—Usual Ratios of Oil, Water and Gum Used to Produce Emulsions

System	Acacia	Tragacanth
Fixed oils (excluding liquid petrolatum and linseed oil)	4	40
Water	2	20
Gum	1	1
Volatile oils, plus liquid petrolatum and linseed oil	2–3	20–30
Water	2	20
Gum	1	1

mulation are separated into those that are oil-soluble and those that are water-soluble. These are dissolved in their respective solvents by heating to about 70 to 75°. When solution is complete, the two phases are mixed and the product is stirred until cool. This method, which requires nothing more than two beakers, a thermometer and a source of heat, is necessarily used in the preparation of emulsions containing waxes and other high-melting-point materials that must be melted before they can be dispersed in the emulsion. The relatively simple methodology involved in the use of synthetic surfactant-type emulsifiers is one factor which has led to their widespread use in emulsion preparation. This, in turn, has led to a decline in the use of the natural emulsifying agents.

With hand homogenizers an initial rough emulsion is formed by trituration in a mortar or shaking in a bottle. The rough emulsion is then passed several times through the homogenizer. A reduction in particle size is achieved as the material is forced through a narrow aperture under pressure. A satisfactory product invariably results from the use of a hand homogenizer and overcomes any deficiencies in technique. Should the homogenizer fail to produce an adequate product, the formulation, rather than the technique, should be suspected.

For a discussion of the techniques and equipment used in the large-scale manufacture of emulsions, see Chapter 83.

Stability of Emulsions

There are several criteria which must be met in a well-formulated emulsion. Probably the most important and most readily apparent requirement is that the emulsion possess adequate physical stability; without this, any emulsion soon will revert back to two separate bulk phases. In addition, if the emulsified product is to have some antimicrobial activity (eg, a medicated lotion), care must be taken to ensure that the formulation possesses the required degree of activity. Frequently, a compound exhibits a lower antimicrobial activity in an emulsion than, say, in a solution. Generally, this is because of partitioning effects between the oil and water phases, which cause a lowering of the "effective" concentration of the active agent. Partitioning has also to be taken into account when considering preservatives to prevent microbiological spoilage of emulsions. Finally, the chemical stability of the various components of the emulsion should receive some attention, since such materials may be more prone to degradation in the emulsified state than when they exist as a bulk phase.

In the present discussion, detailed consideration will be limited to the question of physical stability. Reviews of this topic have been published by Garrett[63] and Kitchener and Mussellwhite.[64] For information on the effect that emulsification can have on the biologic activity and chemical stability of materials in emulsions, see Wedderburn,[65] Burt[66] and Swarbrick.[67]

The theories of emulsion stability have been discussed by Eccleston[68] in an attempt to understand the situation in both a simple O/W emulsion and complex commercial systems.

The three major phenomena associated with physical stability are

1. The upward or downward movement of dispersed droplets relative to the continuous phase, termed *creaming* or *sedimentation*, respectively
2. The aggregation and possible coalescence of the dispersed droplets to reform the separate, bulk phases.
3. Inversion, in which an O/W emulsion inverts to become a W/O emulsion, and *vice versa*.

Creaming and Sedimentation—Creaming is the upward movement of dispersed droplets relative to the continuous phase, while sedimentation, the reverse process, is the downward movement of particles. In any emulsion one process or the other takes place, depending on the densities of the disperse and continuous phases. This is undesirable in a pharmaceutical product where homogeneity is essential for the administration of the correct and uniform dose. Furthermore, creaming, or sedimentation, brings the particles closer together and may facilitate the more serious problem of coalescence.

The rate at which a spherical droplet or particle sediments in a liquid is governed by Stokes' law (Eq 35). While other equations have been developed for bulk systems, Stokes' equation is still useful since it points out the factors that influence the rate of sedimentation or creaming. These are the diameter of the suspended droplets, the viscosity of the suspending medium, and the difference in densities between the dispersed phase and the dispersion medium.

Usually, only the use of the first two factors is feasible in affecting creaming or sedimentation. Reduction of particle size contributes greatly toward overcoming or minimizing creaming, since the rate of movement is a square-root function of the particle diameter. There are, however, technical difficulties in reducing the diameter of droplets to below about 0.1 μm. The most frequently used approach is to raise the viscosity of the continuous phase, although this can be done only to the extent that the emulsion still can be removed readily from its container and spread or administered conveniently.

Aggregation and Coalescence—Even though creaming and sedimentation are undesirable, they do not necessarily result in the breakdown of the emulsion, since the dispersed droplets retain their individuality. Furthermore, the droplets can be redispersed with mild agitation. More serious to the stability of an emulsion are the processes of aggregation and coalescence. In aggregation (flocculation) the dispersed droplets come together but do not fuse. Coalescence, the complete fusion of droplets, leads to a decrease in the number of droplets and the ultimate separation of the two immiscible phases. Aggregation precedes coalescence in emulsions; however, coalescence does not necessarily follow from aggregation. Aggregation is, to some extent, reversible. While not as serious as coalescence, it will accelerate creaming or sedimentation, since the aggregate behaves as a single drop.

While aggregation is related to the electrical potential on the droplets, coalescence depends on the structural properties of the interfacial film. In an emulsion stabilized with surfactant-type emulsifiers forming monomolecular films, coalescence is opposed by the elasticity and cohesiveness of the films sandwiched between the two droplets. In spite of the fact that two droplets may be touching, they will not fuse until the interposed films thin out and eventually rupture. Multilayer and solid-particle films confer on the emulsion a high degree of resistance to coalescence, due to their mechanical strength.

Particle-size analysis can reveal the tendency of an emulsion to aggregate and coalesce long before any visible signs of instability are apparent. The methods available have been reviewed by Groves and Freshwater.[69]

Inversion—An emulsion is said to invert when it changes from an O/W to a W/O emulsion, or *vice versa*. Inversion sometimes can be brought about by the addition of an electrolyte or by changing the phase-volume ratio. For example, an O/W emulsion having sodium stearate as the emulsifier can be inverted by the addition of calcium chloride, because the calcium stearate formed is a lipophilic emulsifier and favors the formation of a W/O product.

Inversion often can be seen when an emulsion, prepared by heating and mixing the two phases, is being cooled. This takes place presumably because of the temperature-depen-

dent changes in the solubilities of the emulsifying agents. The phase inversion temperature, or PIT, of nonionic surfactants has been shown by Shinoda, *et al*[70] to be influenced by the HLB number of the surfactant. The higher the PIT value, the greater the resistance to inversion.

Apart from work on PIT values, little quantitative work has been carried out on the process of inversion; nevertheless, it would appear that the effect can be minimized by using the proper emulsifying agent in an adequate concentration. Wherever possible, the volume of the dispersed phase should not exceed 50% of the total volume of the emulsion.

Bioavailability from Coarse Dispersions

In recent years, considerable interest has focused on the ability of a dosage form to release drug following administration to the patient. Both the rate and extent of release are important. Ideally, the extent of release should approach 100%, while the rate of release should reflect the desired properties of the dosage form. For example, with products designed to have a rapid onset of activity, the release of drug should be immediate. With a long-acting product, the release should take place over several hours, or days, depending on the type of product used. The rate and extent of drug release should be reproducible from batch to batch of the product, and should not change during shelf life.

The principles on which biopharmaceutics is based are dealt with in some detail in Chapters 35 to 37. While most published work in this area has been concerned with the bioavailability of solid dosage forms administered by the oral route, the rate and extent of release from both suspensions and emulsions is important and so will be considered in some detail.

Bioavailability from Suspensions—Suspensions of a drug may be expected to demonstrate improved bioavailability compared to the same drug formulated as a tablet or capsule. This is because the suspension already contains discrete drug particles, whereas tablet dosage forms must invariably undergo disintegration in order to maximize the necessary dissolution process. Frequently, antacid suspensions are perceived as being more rapid in action and therefore more effective than an equivalent dose in the form of tablets. Bates, *et al*[71] observed that a suspension of salicylamide was more rapidly bioavailable, at least during the first hour following administration, than two different tablet forms of the drug; these workers were also able to demonstrate a correlation between the initial *in vitro* dissolution rates for the several dosage forms studied and the initial rates of *in vivo* absorption. A similar argument can be developed for hard gelatin capsules, where the shell must rupture or dissolve before drug particles are released and can begin the dissolution process. Such was observed by Antal, *et al*[72] in a study of the bioavailability of several doxycycline products, including a suspension and hard gelatin capsules. Sansom, *et al*[73] found mean plasma phenytoin levels higher after the administration of a suspension than when an equivalent dose was given as either tablets or capsules. It was suggested that this might have been due to the suspension having a smaller particle size.

In common with other products in which the drug is present in the form of solid particles, the rate of dissolution and thus potentially the bioavailability of the drug in a suspension can be affected by such factors as particle size and shape, surface characteristics, and polymorphism. Strum, *et al*[74] conducted a comparative bioavailability study involving two commercial brands of sulfamethiazole suspension (Product A and Product B). Following administration of the products to 12 normal subjects and taking blood samples at predetermined times over a period of 10 hr, the workers found no statistically significant difference in the extent of drug absorption from the two suspensions. The absorption rate, however, differed, and from *in vitro* studies it was concluded that product A dissolved faster than product B and that the former contained more particles of smaller size than the latter, differences that may be responsible for the more rapid dissolution of particles in product A. Product A also provided higher serum levels in *in vivo* tests half an hour after administration. The results showed that the rate of absorption of sulfamethiazole from a suspension depended on the rate of dissolution of the suspended particles, which in turn was related to particle size. Previous studies[75,76] have shown the need to determine the dissolution rate of suspensions in order to gain information as to the bioavailability of drugs from this type of dosage form.

The viscosity of the vehicle used to suspend the particles has been found to have an effect on the rate of absorption of nitrofurantoin but not the total bioavailability. Thus Soci and Parrott were able to maintain a clinically acceptable urinary nitrofurantoin concentration for an additional two hours by increasing the viscosity of the vehicle.[77]

Bioavailability from Emulsions—There are indications that improved bioavailability may result when a poorly absorbed drug is formulated as an orally administered emulsion. However, little study appears to have been made in direct comparison of emulsions and other dosage forms such as suspensions, tablets, and capsules; thus it is not possible to draw unequivocal conclusions as to advantages of emulsions. If a drug with low aqueous solubility can be formulated so as to be in solution in the oil phase of an emulsion, its bioavailability may be enhanced. It must be recognized, however, that the drug in such a system has several barriers to pass before it arrives at the mucosal surface of the gastrointestinal tract. For example, with an oil-in-water emulsion, the drug must diffuse through the oil globule and then pass across the oil/water interface. This may be a difficult process, depending on the characteristics of the interfacial film formed by the emulsifying agent. In spite of this potential drawback, Wagner, *et al*[78] found that indoxole, a nonsteroidal anti-inflammatory agent, was significantly more bioavailable in an oil-in-water emulsion than in either a suspension or a hard gelatin capsule. Bates and Sequeira[79] found significant increases in maximum plasma levels and total bioavailability of micronized griseofulvin when formulated in a corn oil/water emulsion. In this case, however, the enhanced effect was not due to emulsification of the drug in the oil phase *per se* but more probably because of the linoleic and oleic acids present having a specifical effect on gastrointestinal motility.

References

1. Semat H: *Fundamentals of Physics*, 3rd ed, Holt-Rinehart-Winston, New York, 1957.
2. Michaels AS: *J Phys Chem 65*. 1730, 1961.
3. Zisman WA: *Adv Chem Ser 43*· 1, 1964.
4. Titoff Z: *Z Phys Chem 74*. 641, 1910.
5. Lowell S: *Introduction to Power Surface Area*, Wiley-Interscience, New York, 1979.
6. Osipow LI: *Surface Chemistry Theory and Industrial Applications*, Reinhold, New York, 1962
7. Langmuir I: *J Am Chem Soc 39* 1848, 1917.
8. Giles CH: In EH Lucassen-Reynders, ed, *Anionic Surfactants*, Marcel Dekker, New York, 1981, Ch 4.
9. Weiser HB: *A Textbook of Colloid Chemistry*, Elsevier, New York, 1949.
10. Ter-Minassian-Saraga L: *Adv Chem Ser 43* 232, 1964.
11. Schott H, Martin AN. In Dittert LW, ed, *American Pharmacy*, 7th ed, Lippincott, Philadelphia, Chap 6, 1974.

12. Shinoda K, Nakagawa T, Tamamushi BI, Isemura T: *Colloidal Surfactants*, Academic, New York, Chap 2, 1963.
13. Sjoblom L. In Shinoda K, ed, *Solvent Properties of Surfactant Solutions*, Marcel Dekker, New York, Chap 5, 1967.
14. Prince, LM: *Microemulsions—Theory and Practice*, Academic, New York 1977.
15. Shinoda K, Friberg S: *Adv Colloid Interface Sci 4*: 281, 1975.
16. Friberg SE, Venable RV in Becher P, ed: *Encyclopedia of Emulsion Technology*, vol 1, Marcel Dekker, Chapt 4, New York, 1983.
17. Overbeek JThG: *Disc Faraday Soc 65*. 7, 1978.
18. Davis SS. In Bundgaard H, Hansen AB, Kofod H, eds: In *Optimization of Drug Delivery*, Alfred Benzon Symposium 17, Munksgaard, Copenhagen, 1982.
19. Fendler JH, Fendler EJ: *Catalysis in Micellar and Macromolecular Systems*, Academic, New York, 1975.
20. Mackay RA: *Adv Colloid Interface Sci 15*, 131, 1981.
21. Kruyt HR: *Colloid Science*, vols I and II, Elsevier, Houston, 1952 and 1949.
22. Alexander AE, Johnson P: *Colloid Science*, Oxford University Press, Oxford, 1949.
23. Ross S, Morrison ID: *Colloidal Systems and Interfaces*, Wiley, New York, 1988.
24. Mysels KJ: *Introduction to Colloid Chemistry*, Wiley-Interscience, New York, 1959
25. Shaw DJ: *Introduction to Colloid and Surface Chemistry*, 3rd ed, Butterworths, London, 1980.
26 Vold RD, Vold MJ: *Colloid and Interface Chemistry*, Addison-Wesley, Reading, MA., 1983.
27. Hiemenz PC. *Principles of Colloid and Surface Chemistry*, Marcel Dekker, New York, 1986.
28. von Weimarn PP. In Alexander J, ed: *Colloid Chemistry*, vol I, Chemical Catalog Co (Reinhold), New York, 1926. See also *Chem Rev 2* 217, 1926.
29. Lachman L *et al* *Theory and Practice of Industrial Pharmacy*, 3rd ed, Lea & Febiger, Philadelphia, 1986
30. Tubis M, Wolf W, eds: *Radiopharmacy*, Wiley-Interscience, New York, 1976.
31. LaMer VK, Dinegar RH. *J Am Chem Soc 72* 4847, 1950.
32 Matijevic E: *Acc Chem Res 14*. 22, 1981 and *Ann Rev Mater Sci 15* 483, 1985.
33 Florence AT, Attwood D: *Physicochemical Principles of Pharmacy*, Chapman & Hall, New York, 1982.
34. Gutch CF, Stoner MH: *Review of Hemodialysis*, Mosby, St. Louis, 1975.
35. Schott H, Martin AN. In Dittert LW, ed: *American Pharmacy*, 7th ed, Lippincott, Philadelphia, 1974.
36. Parks GA: *Chem Rev 65* 177, 1965.
37. Schott H: *J Pharm Sci 66*: 1548, 1977.
38. Sonntag H, Strenge K: *Coagulation and Stability of Disperse Systems*, Halstead, New York, 1972.
39. Hough DB, Thompson L. In Schick MJ, ed: *Nonionic Surfactants-Physical Chemistry*, 2nd ed, Marcel Dekker, Chapt 11, New York, 1987.
40. Vincent B: *Adv Colloid Interface Sci 4* 193, 1974.
41. Hunter RJ: *Zeta Potential in Colloid Science*, Academic, New York, 1981.
42. Davies JT, Rideal EK: *Interfacial Phenomena*, 2nd ed, Academic, New York, 1963
43. Bier M, ed: *Electrophoresis*, vols I and II, Academic, New York, 1959 and 1967.
44. Shaw DJ: *Electrophoresis*, Academic, New York, 1969.
45. Cawley LP. *Electrophoresis and Immunoelectrophoresis*, Little-Brown, Boston, 1969.
46. Catsimpoolas N, ed: *Isoelectric Focusing and Isotachophoresis*, *Ann NY Acad Sci 209*. June 15, 1973.
47. Morawetz H: *Macromolecules in Solution*, 2nd ed, Wiley-Interscience, New York, 1975.
48 Veis A: *The Macromolecular Chemistry of Gelatin*, Academic, New York, 1964.
49. Ward AG, Courts A, eds: *The Science and Technology of Gelatin*, Academic, Chapt 6, New York, 1977.
50. Haines BA, Martin A: *J Pharm Sci 50* 228, 753, 756, 1961.
51. Matthews BA, Rhodes CT: *J Pharm Pharmacol 20* Suppl: 204S, 1968.
52. Matthews BA, Rhodes CT: *J Pharm Sci 57* 569, 1968.
53. *Ibid 59* 521, 1970.
54. Schneider W, *et al*: *Am J Pharm Ed 42*: 280, 1978.
55. Tingstad J *et al*: *J Pharm Sci 62*: 1361, 1973.
56. Davies JT: *Proc Intern Congr Surface Activity*, 2nd, London, 426, 1957.
57. Sherman P: In *Emulsion Science*, Chap 4, Academic, New York, 1968.
58. Rogers JA: *Cosmet Toiletries 93*: 49, July, 1978.
59. Griffin WC: *J Soc Cos Chem 1*: 311, 1949.
60. Griffin WC: *J Soc Cos Chem 5*: 249, 1954.
61. Davies JT, Rideal EK: *Interfacial Phenomena*, Chap 8, Academic, New York, 1961. Davies JT: *Proc Intern Congr Surface Activity*, 2nd, London, 426, 1957.
62. Schott J: *J Pharm Sci 60*: 649, 1971.
63. Garrett ER: *J Pharm Sci 54* 1557, 1965.
64. Kitchener JA, Mussellwhite PR. In *Emulsion Science*, Academic, New York, Chap 2, 1968.
65. Wedderburn DL. In *Advances in Pharmaceutical Sciences*, vol 1, Academic, London, 195, 1964.
66. Burt BW: *J Soc Cosm Chem 16*: 465, 1965.
67. Swarbrick J: *Ibid 19*: 187, 1968.
68. Eccleston GM: *Cosmet Toiletries*: 101, 73, Nov 1986.
69. Groves MJ, Freshwater DC: *J Pharm Sci 57*: 1273, 1968.
70. Shinoda K, Kunieda H: In *Encyclopedia of Emulsion Technology*, Chap 5, Marcel Dekker, New York, 1983.
71. Bates TR *et al*. *J Pharm Sci 58* 1468, 1969
72. Antal EJ *et al*: *Ibid 64*. 2015, 1975.
73. Sansom LN *et al* *Med J Aust* 1975(2): 593.
74. Strum JD *et al*. *J Pharm Sci 67*: 1659, 1978.
75. Bates TR *et al*: *Ibid 62*. 2057, 1973.
76. Howard SA *et al* *Ibid 66*: 557, 1977.
77. Soci MM, Parrott EL: *Ibid 69*: 403, 1980.
78. Wagner JG *et al*: *Clin Pharmacol Ther 7*. 610, 1966.
79. Bates TR, Sequeira JA: *J Pharm Sci 64*: 793, 1975.

Bibliography

Interfacial Phenomena

Adamson AW: *Physical Chemistry of Surfaces*, 4th ed, Interscience, New York, 1982.

Davis JT, Rideal EK: *Interfacial Phenomena*, 2nd ed, Academic, New York, 1963.

Hiemenz, PC: *Principles of Colloid and Surface Chemistry*, 2nd ed, Marcel Dekker, New York, 1986.

Shaw DJ: *Introduction to Colloid and Surface Chemistry*, Butterworths, London, 1980.

Colloidal Dispersions

Particle Phenomena and Coarse Dispersions

Davies JT, Rideal EK: *Interfacial Phenomena*, Academic, New York, 1963.

Osipow LI: *Surface Chemistry*, Reinhold, New York, 1962.

Hiemenz PC: *Principles of Colloidal and Surface Chemistry*, Marcel Dekker, New York, 2nd ed, 1986.

Matijevic E, ed: *Surface and Colloid Science*, vols 1–4, Wiley, New York, 1971.

Cadle RD: *Particle Size*, Reinhold, New York, 1965.

Parfitt G: *Dispersion of Powders in Liquids*, Applied Science, 1973.

Adamson AW: *Physical Chemistry of Surfaces*, 4th ed, Wiley-Interscience, New York, 1980.

Fowkes FM, ed: *Hydrophobic Surfaces*, Academic, New York, 1969

Sherman P: *Rheology of Emulsions*, Macmillan, New York, 1963

Becher P: *Emulsions Theory and Practice*, 2nd ed, Reinhold, New York, 1965.

Vold RD, Vold MJ, *Colloid and Interface Chemistry*, Addison-Wesley, Reading, Mass, 1983.

Becher P: *Encyclopedia of Emulsion Technology*, Vols 1 to 3, Marcel Dekker, New York, 1983–1988.

CHAPTER 28

Clinical Analysis

Robert D Smyth, PhD
Vice President, Pharmaceutical Development

Lorraine Evans, BS, H(ASCP)
Clinical Pathology
Pharmaceutical Research & Development Div
Bristol-Myers Company
Syracuse, NY 13221

The characterization and quantitation of the various components of blood, urine and other body fluids are the primary functions of the clinical laboratory. The major divisions of clinical analysis are clinical biochemistry, hematology, blood-bank technology, histopathology, immunology and microbiology. The accurate diagnosis of disease and determination of a potential therapeutic regimen frequently are based on the laboratory analysis of blood, urine, feces, gastric secretions or cerebrospinal fluid. Modern medical practice is tending toward greater reliance on laboratory results as definitive measures of pathological or normal states.

The pharmacist should familiarize himself with the basic principles involved in sample collection, analysis and diagnostic significance of the various clinical parameters. His role in community health necessitates his comprehension of the methodology and diagnostic value of clinical laboratory procedures. The influence of various drugs and drug interactions on these parameters must be considered in both the clinical and drug-abuse situation.

Hematology

The determination of the morphological, physiological and biochemical properties of peripheral blood and the blood-forming organs (hematopoietic system) is a function of the hematology laboratory. The functional categories of hematology are (1) analysis of cellular elements, and specific biochemical and physiological parameters of peripheral blood and the hematopoietic system, (2) blood-coagulation analysis and (3) blood-bank technology.

Peripheral blood is a biphasic liquid tissue system of cellular elements suspended in a liquid plasma phase. The cellular phase comprises about 45% of the blood volume and contains erythrocytes (red blood cells, RBC), leukocytes (white blood cells, WBC) and thrombocytes (platelets). The plasma phase is primarily water (90 to 92%) and protein (7%).

The hematological analysis of blood is concerned primarily with enumeration and differentiation of the various cellular elements. An analysis of the hematopoietic system (eg, bone marrow and lymphoid tissue) determines the status of blood-cell precursors in these tissues. Determinations of specific biochemical (hemoglobin) and physiological (blood or plasma volume) parameters are performed in a complete evaluation of the erythron system (blood and marrow RBC and their precursors). The normal hematological values in the adult are presented in Table I.

The authors acknowledge the assistance of Dr Joseph P Uscavage of Rorer Group Ltd, in the preparation of the *Microbiology* section and Dr Alfred H Free of the Ames Co for the *Urinalysis* section.

Erythrocytes and Hemoglobin—The erythrocytic system is composed of the mature erythrocytes in peripheral blood and their precursors in bone marrow. The precursors of erythrocytes, as found in the erythropoietic system (red bone marrow), are classified as to the degree of nucleation and characteristics of cytoplasmic constituents. The sequence of erythrocyte formation in bone marrow—based on the gradual denucleation of the cell, generation of the chromatin structure and changes in nucleolar structure and cytoplasmic constituents—is as follows:

pronormoblast → basophilic normoblast → polychromatic normoblast → orthochromatic normoblast → polychromatophilic erythrocyte → erythrocyte.

The first four types are nucleated and normally are seen only in bone marrow. In normal erythrocyte formation these immature bone-marrow cells are designated as *normoblastic* or *normocytic*. In pernicious anemia and related conditions they become abnormally large and are designated *megaloblastic* or *megalocytic*. In iron-deficiency anemia, these cells become abnormally small and are designated *microblastic* or *microcytic*—of the iron-deficiency type.

Table I—Normal Hematological Values in Man[1]

	Normal Value	Normal Range of Values
Erythrocytes (cu mm × 10^6)		
Male	5.4	4.6–6.2
Female	4.8	4.2–5.6
Reticulocytes (cu mm × 10^3)	50	10–100
Hemoglobin (g%)		
Male	16.0	14.0–18.0
Female	14.0	12.0–16.0
Hematocrit (%)		
Male	47.0	40.0–54.0
Female	42.0	37.0–47.0
Mean corpuscular volume (μm^3)	87	82–92
Mean corpuscular hemoglobin (pg)	29	27–31
Mean corpuscular hemoglobin concentration (%)	34	32–36
Mean corpuscular diameter (μm)	7.3	6.7–7.7
Leukocytes (cu mm × 10^3)	7.0	5.0–10.0
Leukocyte differential (%)		
Neutrophils	63	57–67
Eosinophils	1	1–3
Basophils	1	0–1
Lymphocytes	30	25–33
Monocytes	5	3–7
Platelets (cu mm × 10^6)	3.0	1.4–6.0
Erthrocyte sedimentation rate (Wintrobe), (mm/hr)		
Male	4	0–9
Female	10	0–20

Normal blood contains 0.5 to 1.5% of circulating erythrocytes as reticulocytes. These cells contain a fine network of basophilic reticulum that is demonstrable on staining with a vital dye such as brilliant cresyl blue. The number of these cells in the blood is a measure of effective erythropoiesis. High circulating-reticulocyte values are an index of erythropoietic activity and are found in the first few days of life, after hemorrhage and after treatment of iron- or vitamin B_{12}-deficiency anemias.

The normal *erythrocyte* (normocyte) is a flexible, elastic, biconcave, enucleated structure with a mean diameter of 7.3 μm and a thickness near 2.2 μm. The chemical constituents of the red blood cell include water (63%), lipids (0.5%), glucose (0.8%), minerals (0.7%), nonhemoglobin protein (0.9%), methemoglobin (0.5%) and hemoglobin (33.6%). The primary function of the erythrocyte is transport of oxygen and carbon dioxide. The red cell membrane, a dynamic, semipermeable component of the cell, is associated with energy metabolism in the maintenance of the permeability characteristics of the cell to various cations (Na^+, K^+) and anions (Cl^-, HCO_3^-). The stroma of insoluble material which remains after red-cell disruption (hemolysis) constitutes 2 to 5% of the wet-cell weight; it is primarily protein (40 to 60%) and lipid (10 to 12%). The membrane includes stromatin (a fibrous or structural protein) and mucopolysaccharides associated with A, B and O blood-group substances. The lipid fractions include phosphatides (lecithin, cephalin), cholesterol, cholesterol esters, neutral fats, cerebrosides and sialic acid glycoproteins.

Erythrocytes may be enumerated by either visual or electronic procedures. In the visual procedures, a measured quantity of blood is diluted with a fluid which is isotonic with blood and will prevent its coagulation. The diluted blood is then placed in a counting chamber (hemocytometer), and the number of cells in a circumscribed area is enumerated microscopically. Hayem's solution (sodium sulfate, 2.5 g; sodium chloride, 0.25 g; mercuric chloride, 0.25 g; distilled water, 100 mL), Toison's fluid (sodium sulfate, 8 g; sodium chloride, 1 g; methyl violet, 0.025 g; glycerin, 30 mL; distilled water, 180 mL) or 0.9% sodium chloride are used as diluting fluids. The overall error of this method is about 8%.

A greater degree of accuracy and reproducibility can be achieved by erythrocyte enumeration in an electronic counting apparatus; eg, Coulter Counter or Ortho cell counters. The Coulter method (Fig 28-1) determines the number and size of particles suspended in an electrically conductive liquid. The blood cells traverse a small aperture and displace their own volume in the diluent as to produce a change in resistance between the electrodes; the magnitude of the voltage pulse is proportional to cell volume, and the resultant pulses are then amplified, scaled and automatically counted.

In the Ortho ELT-8 technique (Fig 28-2), the principles of laser flow cytometry are used to count cells. Hydrodynamic focusing and laminar flow are combined in the system to count a large number of individual cells. Light focused by a helium-neon laser is scattered by the cells as they pass through the flow channel. The scattered light is monitored by a photoelectric sensor and transfers the electrical pulses which are processed by the systems circuitry. In addition to increased counting speed, the overall error of the electronic procedures is reduced to about 1%.

The *hematocrit value* is also a measure of the erythrocyte portion of blood. A sample of blood containing an anticoagulant is placed in a graduated hematocrit capillary tube, centrifuged and the volume ratio of packed red cells to total blood volume (hematocrit value) determined. The centrifuged sample appears as a red layer of packed erythrocytes over which is found an off-white layer of packed leukocytes and platelets, and a supernatant plasma phase. The hematocrit value is an index of both the number and size of the red cells.

Hemoglobin, a conjugated hemoprotein with an approximate molecular weight of 67,000, contains basic proteins, the globins and ferroprotoporphyrin (heme). It is essentially a tetramer, consisting of four peptide chains, to each of which is bound a heme group. Heme, which constitutes about 4% of the weight of the molecule, consists of a divalent iron atom in the center of a pyrrole-porphyrin structure. Four distinct polypeptide chains (α, β, γ, δ) can be incorporated into hemoglobin. Normal adult hemoglobin is $HbA = \alpha_2^A\beta_2^A$. Fetal hemoglobin contains 2α and 2γ chains and is designated $HbF = \alpha_2^A\gamma_2^F$.

Differences in the structural sequences of amino acids in the peptide portion of the hemoglobin molecules are controlled genetically and are responsible for different types of hemoglobin. Based on the characteristic mobility of the hemoglobin, in an electric field (electrophoresis) on starch, paper, cellulose acetate, agar or acrylamide gel media, many hemoglobin types have been recognized (see Chapter 29). Only types P, F and A_1-A_4 are considered normal. Sickle-cell anemia and β-thalassemia are hemolytic anemias associated with abnormal hemoglobins (ie, Type S in sickle-cell anemia and abnormal production of the β chain in β-thalas-

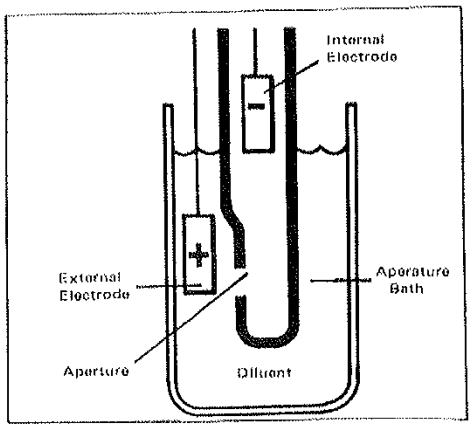


Fig 28-1. Coulter-counting cells by electronic impedance (courtesy Coulter Electronics).

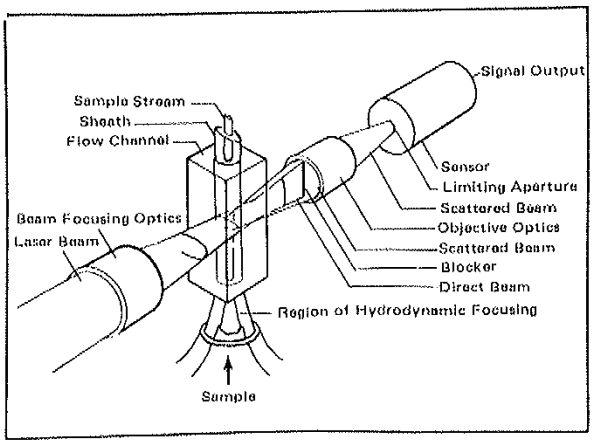


Fig 28-2. Ortho ELT-8-Method of scattered light detection and hydrodynamic focusing for cell counting (courtesy, Clinical Instrument Systems, Oct 1980).

semia). In homozygous *HbS disease* sickling of the red cells is due to the low solubility of the abnormal hemoglobin in its reduced state, with the production of semicrystalline bodies (tactoids), which distort and elongate the cells. In the sickle-cell trait (heterozygous), the blood smear shows no sickle cells. In the homozygous condition, HbS accounts for nearly all of the hemoglobin with small amounts of HbF. In the heterozygous condition, HbS constitutes 50% or less of the hemoglobin, with the balance as HbA.

The detection of sickle-cell disease is performed by microscopic observation of the induction of red-cell sickling in the presence of a reducing agent such as sodium metabisulfite or by quantitative determination of urea-dispersible turbidity induced by dithionite following reduction of HbS to deoxy-HbS in RBC lysates. The microscopic procedure will detect only homozygotes, whereas HbAS and HbS and its structural variant HbC-Harlem both are detected in the urea-dithionite technique. Commercial qualitative test kits are available for detecting sickle-cell trait and anemia by solubility determinations. All hemoglobins positive to the dithionite test must be electrophorized (cellulose acetate, citrate agar or starch gel) to differentiate HbS from HbC and thalassemia traits. Drugs causing hemolysis in glucose 6-phosphate dehydrogenase (G6PD) deficiency include sulfones, nitrofurans, chloroquine, dimercaprol, nalidixic acid and probenecid.

The *hemoglobin* concentration is measured spectrophotometrically after lysis of whole blood and conversion of hemoglobin to hematin, oxyhemoglobin or cyanmethemoglobin. The addition of a strong base (NaOH) to pH 10 converts oxyhemoglobin, carboxyhemoglobin and methemoglobin to hematin, which can be estimated photometrically. Weaker bases (Na_2CO_3 or NH_4OH) convert hemoglobin to oxyhemoglobin for analysis.

Total hemoglobin is measured also by conversion to cyanmethemoglobin using alkaline sodium cyanide-potassium ferricyanide reagent. Hemoglobin standards certified by the Clinical Standards Committee of the College of American Pathologists are used in these procedures, and all results are expressed as "g hemoglobin per 100 mL blood."

In the normal state, the oxygen consumption of the RBC is low and it is involved in the conversion of hemoglobin to oxidized (Fe^{3+}) methemoglobin (HbM) which cannot bind oxygen. The normal balance of HbM (<0.5%) is maintained by two enzyme systems—NADH and NADPH methemoglobin reductases. An inherited deficiency of the RBC enzyme, G6PD. This will decrease the rate of reduction of glutathione and methemoglobin, make the cell more vulnerable to oxidative attack and result in susceptibility to drug-induced or immune-mediated nonspherocytic hemolytic anemia. G6PD deficiency is found predominantly in Mediterranean peoples, Southeast Asians, Africans and American negroes. The enzyme can be quantitated spectrometrically or by fluoronephelometry by measuring the rate of reduction of nicotinamide adenine dinucleotide phosphate (NADP) in the presence of G6PD. Presumptive screening tests based on reduced glutathione (GSH) content of blood before and after incubation with acetylphenylhydrazine also are used.

Erythrocyte count, hemoglobin content and hematocrit value are used to determine various blood indices in the diagnosis and treatment of anemia. These measurements are:

Mean corpuscular volume {MCV (μm^3)} =

$$\frac{\text{Hematocrit (\%)} \times 10}{\text{Erythrocyte count (millions/cu mm)}}$$

Mean corpuscular hemoglobin {MCH (pg)} =

$$\frac{\text{Hemoglobin (g/100 mL)} \times 10}{\text{Erythrocyte count (millions/cu mm)}}$$

Mean corpuscular hemoglobin concentration {MCHC(%)} =

$$\frac{\text{Hemoglobin (g/100 mL)} \times 100}{\text{Hematocrit (\%)}}$$

An additional parameter used to characterize red-cell variation is the red-cell distribution width (RDW) determined on the Coulter S-Plus II. The RDW is calculated directly by the standard deviation and coefficient of variation from a red-cell histogram on the S-Plus II. The difference in cell size may be used to monitor patients with pernicious or hemorrhagic anemia.

Anemias are classified as to red-cell volume and hemoglobin concentration. *Macrocytic* (large cell: MCV > 94), *normocytic* (normal cell: MCV, 82 to 92), or *microcytic* (small cell: MCV < 80) are the classifications according to cell volume. Cellular hemoglobin concentration categorizes the cells as to *hyperchromic* (MCHC > 38), *normochromic* (MCHC = 32 to 36), or *hypochromic* (MCHC < 30). Examples of anemias:

I. Hypochromic Microcytic—erythroid normoblastic anemia in bone marrow
   - A. Iron Deficiency—low hemoglobin (Hbg) and RBC, low serum iron, high total iron binding capacity, absent hemosiderin.
     1. Dietary—low iron intake
     2. Intestinal problems—decreased iron absorption
     3. Pregnancy, infants—increased iron requirements
     4. Iron loss—due to chronic hemorrhage, parasitic infections, GI tract lesions, excess menstrual bleeding.
   - B. Hereditary Sideroblastic—defect in the heme synthesis, an inability to utilize ingested iron.
   - C. Thalassemia—genetic abnormality which produces normal to increased HbgF and/or $HbgA_2$.

II. Normochromic Normocytic
   - A. Hemolytic—increased destruction of erythrocytes.
     1. Autoimmune hemolytic
     2. Cold agglutinin hemolytic
     3. Mechanical destruction of RBCs
     4. Paroxysmal Nocturnal hemoglobinuria
     5. Lymphomas and Hodgkin's disease
     6. Infections
   - B. Hemoglobinopathies—abnormalities in structure of alpha or beta chains of hemoglobin molecule; normoblastic erythroidhyperplasia in bone marrow.
     1. Sickle-cell
     2. Hemolysis
     3. Hemoglobin CC
   - C. Acute Hemorrhage
   - D. Other
     1. Aplastic Anemia, Leukemia, Malignancy
     2. Renal failure and drug-related anemias caused by chloramphenicol and antineoplastic drugs.

III. Normochromic Macrocytic—due to deficiency of vitamin B_{12} or folate; bone marrow is hypercellular with increased erythroid precursors.
   1. Pernicious
   2. Sideroblastic
   3. Sprue—total iron-binding capacity is decreased; hemosiderin is increased in the bone marrow.
   4. Pregnancy

Determinations of the suspension stability of whole blood and erythrocyte fragility are useful adjuncts in the diagnosis of various diseases.

The *erythrocyte sedimentation rate* (ESR) is an estimate of the suspension stability of red blood cells in plasma; it is related to the number and size of the red cells and to the relative concentration of plasma proteins, especially fibrinogen and the α- and β-globulins. This test is performed by determining the rate of sedimentation of blood cells in a standard tube. Normal blood ESR is 0 to 15 mm/hour. Increases are an indication of active but obscure disease processes such as tuberculosis and ankylosing spondylitis. ESR is affected by anemia and does not respond linearly with changes in asymmetrical macromolecules such as fibrinogen and globins.

The *zeta sedimentation ratio* (ZSR) technique overcomes these disadvantages. It is based on a measure of the closeness with which RBC will

approach each other after standardized cycles of dispersion and compaction.

The *erythrocyte fragility test* is based on resistance of cells to hemolysis in decreasing concentrations of hypotonic saline.

Increased osmotic fragility of the red cells is associated with various types of spherocytosis and acquired hemolytic anemia; increased resistance has been observed in thalassemia, sickle-cell anemia and hypochromic anemia. The test can be performed manually by colorimetric estimation of hemoglobin released by hypotonic cell rupture or automatically in an instrument which continually records the increase in light transmittance through a suspension of red cells in a continuously decreasing salt gradient during dialysis.

Leukocytes—Mature *leukocytes* (white blood cells, WBC) in peripheral blood and their precursors in bone and lymphoid tissue comprise the leukocytic system. Various types of leukocytes are found in normal blood. Differentiation of the lymphocytic, monocytic and granulocytic leukocyte types is based on cell size, color, chromatin structure and cytoplasm constituents.

The primary function of leukocytes is the development of the various defensive and reparative processes in inflammatory and immune-response mechanisms. The migration of leukocytes to the site of inflammation is associated with the release or activation of various biochemical substances (5-hydroxytryptamine, histamine, complement, immunoglobulins, prostaglandins, lysosomal enzymes). The tissue histiocyte or monocyte (macrophage) also can engulf and destroy foreign particles by the process of endocytosis and certain leukocyte types by phagocytosis.

The chemical composition of the leukocyte includes water (82%), nucleoprotein, phospholipids and trace minerals. Enzyme content, glycogen and histamine levels vary in the different types of white cells. Deficiency in enzymes associated with glycolytic metabolism (hexokinase) and increases in phosphomonoester hydrolases (alkaline phosphatase) have been observed in leukocytes of certain leukemia patients.

The precursors of granulocytic leukocytes are found in bone marrow and are classified according to the degree of cytoplasmic granulation, dye-affinity of the granules and shape of the nucleus (Schilling, Arneth or Cooke-Ponder Classification). As undifferentiated cells (myeloblasts) mature

promyelocyte → myelocyte → metamyelocyte → band leukocyte → segmented leukocyte

metachromatic granules appear in the cytoplasm (granulocytes). All segmented leukocytes are motile, a requirement for participation in the inflammatory or phagocytic processes.

In the mature *basophilic* and *eosinophilic leukocytes*, these granules develop an affinity for a basic or acidic dye, respectively; those cells containing granules which do not stain are called *neutrophils*. In peripheral blood, the mature granulocytic cells are designated *polymorphonuclear leukocytes—neutrophilic, eosinophilic or basophilic*.

The other types of white cells normally observed in peripheral blood have no granules and are classified as to size and shape into the *monocyte* and *lymphocyte*, which are formed in lymphoid tissue. The small lymphocyte is thymic-derived and is found in the circulation and germinal centers of lymphoid tissue. The origin of the large lymphocyte is a gut-associated lymphoid stem cell which can further differentiate into the immunoglobulin-producing plasmacyte. The interaction of thymic (T) and bone-marrow (B) lymphocytes is the basis for the development and maintenance of humoral and cellular immune mechanisms.

Leukocytes are enumerated by procedures similar to those used for erythrocytes. In the visual procedures the blood is diluted with a fluid (3% v/v acetic acid) which lyses the red cells, and the total leukocyte count is determined microscopically. Eosinophils also may be analyzed differentially with a diluting fluid which renders the red cells nonrefractile and invisible, and lyses the base-labile leukocytes, leaving the base-stable eosinophils intact. A suitable diluting fluid for this purpose is Pilot's Fluid (propylene glycol, 50 mL; distilled water, 40 mL; 1% phloxine, 10 mL; 10% sodium carbonate, 1 mL and heparin sodium, 100 units). Electronic-counting procedures are similar to those used for erythrocytes with the added advantages of speed, accuracy and reproducibility.

The normal adult leukocyte value is 5000 to 10,000 cells/cu mm. Values greater than 10,000 (*leukocytosis*) are encountered in the newborn infant, young children, after violent exercise, convulsive seizures of epilepsy, leukemia and cancer. Values of less than 5000 (*leukopenia*) are observed in certain microbial infections (eg, typhoid fever, measles, malaria, overwhelming septicemia), cirrhosis of the liver, pernicious anemia, radiation injury and replacement of marrow by malignant tissue.

A *differential count of the leukocytes* provides information as to the relative numbers of each type. A thin film of blood is prepared on a microscope slide stained with a polychromatic preparation such as the Leishman, Wright or Giemsa stain, and analyzed microscopically. Wright's stain contains polychromed methylene blue and eosin dyes; the erythrocytes are stained pink; the nuclei of the leukocytes, purplish-blue; neutrophilic granules, violet-pink; eosinophilic granules, red; basophilic granules, blue; and platelets, blue.

The recent introduction of automated systems for differential white-cell counts significantly reduce the errors inherent with the subjective nature of the visual counting procedure. Differentiation of the various cell types can be made on the basis of cytochemistry and staining properties of enzymes specific for a single cell type. The granules of neutrophils and eosinophils are stained by action of their peroxidases on 4-chloro-1-naphthol to form a colored quinone in the presence of a peroxide and further differentiated by the optimum pH for peroxidase activity between these two cell types. The monocytic lipase is used as a specific marker by the reaction of basic fuchsin with α-naphthol liberated by lipase on α-naphthylbutyrate substrate. The lymphocytes are not stained in this procedure but are measured by electronic sizing.

Automated differential WBC counts also have been obtained in systems which count large populations of cells by simultaneous measurement of two optical properties (axial light loss and/or narrow-angle scatter and/or multiple-wavelength fluorescence). Laser light also is used to differentiate cell size, granularity and volume of cells. The collected light measured by forward versus right-angle scatter is converted to a histogram giving the percent of lymphocytes, monocytes and granulocytes. Another system involves computer processing of two-dimensional images of the various cell types using an automatic scanning microscope.

Polymorphonuclear neutrophilic leukocytes (neutrophils, "polys") normally comprise 62% (50 to 67%) of the total leukocyte count. These cells are irregular in shape (10 to 15 μm in diameter) and usually contain a multilobated nucleus with fine, lightly stained cytoplasmic granules. An immature or juvenile form of neutrophil, with a band-shaped nonsegmented nucleus constitutes 3 to 5% of peripheral blood leukocytes. Increases in the relative percentage of these cells (neutrophilia) is observed in acute microbial infections (eg, meningitis, smallpox, poliomyelitis), metabolic disorders (diabetic acidosis, gout), drug intoxication (digitalis, epinephrine), vaccination, coronary thrombosis and malignant neoplasms.[2]

Polymorphonuclear eosinophilic leukocytes (eosinophils) normally comprise about 1 to 3% of total circulating white-blood cells. In appearance they are similar to the neutrophil with the exception of large, red-stained cytoplasmic granules. Eosinophilia has been observed in certain

skin diseases (psoriasis, eczema), parasitic infestations (pork round worm—trichinosis), certain hypersensitivity reactions, scarlet fever and pernicious anemia. Charcot-Leyden crystals, which are found in bronchial secretions from asthmatics, are derived from nucleoprotein-disintegration products of eosinophils.

Polymorphonuclear basophilic leukocytes (basophils) possess large cytoplasmic granules which stain a deep blue. These cells, which are primarily sources of blood heparin and histamine, constitute less than 1.0% of the leukocytes. Basophilic leukocytosis is seen in chronic myelocytic leukemia, hemolytic anemia and Hodgkin's disease. Basophilic leukopenia occurs following radiation or therapy with glucocorticoids.

Lymphocytes have a cell diameter from 7 to 10 μm (small) to 10 to 18 μm (large). They have a round, or slightly indented, deeply stained nucleus and normally comprise 25 to 33% of the leukocytes. Lymphocytosis is seen in infectious mononucleosis, lymphocytic leukemia, rickets and in most conditions associated with neutrophilic leukopenia (neutropenia).

Monocytes constitute 3 to 7% of the leukocytes. They are larger (12 to 20 μm) than the other leukocytes and possess an abundant, pale, bluish-violet-stained cytoplasm with a fine, reticulated chromatin structure in the nucleus. The monocytes (macrophages) phagocytize bacteria, parasitic protozoa, foreign particles and even erythrocytes. Monocytosis is seen in certain microbial infections (tuberculosis, typhus, malaria), Hodgkin's disease and monocytic leukemia.

Drug therapy frequently causes neutrophil dysfunction which can be characterized by a decreased number of mature neutrophils or a defect in cellular function resulting in the inability of the body to defend itself against infection. Drugs such as nitrogen mustard and chloramphenicol degenerate bone-marrow stem cells, and DNA synthesis is impaired by antimetabolites such as methotrexate and fluorouracil. Depolymerization of DNA is caused by procarbazine and alkylating agents. Mitosis is inhibited by colchicine and vinca alkaloids. The following outline lists drugs which cause granulocytopenia.[2]

- Nonchemotherapeutic
  - rifampin
  - fistocetin
  - benzene
  - nitrous oxide
  - ethanol
- Antithyroid
  - carbimazole
  - methimazole
  - thiouracil
- Diuretics
  - acetazolamide
  - chlorthalidone
  - chlorothiazide
  - ethacrynic acid
  - hydrochlorothiazide
  - mercurials
- Antihistamines
  - ethylenediamine
  - thenalidine
  - metaphenylene
  - pyribenzamine
- Phenothiazines
  - chlorpromazine
  - mepazine
  - methotrimeprazine
  - prochlorperazine
  - thoridizine
- Antibiotics
  - chloramphenicol
  - carbenicillin
  - griseofulvin
  - isoniazid
  - novobiocin
- Cardiovascular
  - diazoxide
  - procainamide
  - methyl dopa
  - quinidine
  - propranolol

As qualitative and quantitative changes in leukocytes in peripheral blood and their precursors in bone marrow and lymphatic tissue are associated with the various types of *leukemia*, this disease has been classified on the basis of the predominating type of leukocyte, ie, myelocytic (granulocytic), lymphocytic, monocytic or plasmacytic. Leukemia may be either acute or chronic and involve the replacement of bone-marrow elements by malignant cells, infiltration of the reticuloendothelial system, anemia, thrombocytopenia and hemorrhage. Leukemia usually is associated with an elevated WBC count and increase in the specific cell and its precursors in peripheral blood, but in certain instances there is an aleukemic blood picture with no evidence of leukocytosis. Leukocytes in acute leukemia are more immature ("blast"-type cells) than those encountered in the chronic type.

In many diseases of the hematopoietic system, it is necessary to examine the bone marrow to determine the rates of formation, maturation and release of blood cells into the peripheral circulation. Using a puncture biopsy needle, samples of *bone marrow* may be obtained from the sternum, iliac crest or proximal end of the tibia. Smears of marrow then are prepared, stained (Wright's stain or specialized histopathological procedure) and examined microscopically. The ratio of myeloid leukocyte to nucleated red cells in bone marrow, the presence of abnormal (*nonmyeloid*) cells, the number of platelet precursors (*megakaryocytes*), the signs of cell-maturation arrest and the presence of focal lesions are important factors in the diagnosis of various disease states.

Systemic lupus erythematosus (SLE) is a disease characterized by numerous clinical and pathological manifestations associated with various organs. Although the disease chiefly affects the lymphatic system, the cardiac, renal and articular systems also are involved. The diagnosis of this disease is based on the presence of an SLE-cell factor in the gamma-globulin fraction of blood in the diseased state. This factor dissolves the nuclei of leukocytes by depolymerization of deoxyribonucleic acid to form the SLE-body. If serum from patients with SLE is incubated with white cells, the "polys" will engulf the liberated SLE-body and form the typical SLE-cell with a characteristic progressive loss of nuclear detail. Drugs which cause SLE and produce a positive SLE-prep include hydralazine, procainamide, isoniazid and phenytoin.

These antibodies to nucleoprotein also can be detected by immunological techniques. In the double-antibody technique, the test serum containing antibodies to nuclear protein is incubated with a rat kidney slice (antigen). The second antibody is a fluorescein-labeled goat antihuman immunoglobulin (IgG) which combines with the human IgG bound to the antigen site in a positive test. The fluorescence is estimated by immunomicroscopy. Normal light-microscopy can be used if the goat-antihuman IgG is labeled with peroxidase.

Thrombocytes—The primary functions of *thrombocytes* (blood platelets) are the maintenance of hemostasis (arrest of blood flow from a vessel) and blood coagulation (clot formation). Platelets are oval to spherical in shape and have a mean diameter of 2 to 4 μm. They originate from an immature cell (megakaryocyte) in bone marrow and ranges of 140,000 to 450,000/cu mm have been reported in normal blood.

Adhesiveness, aggregation and agglutination are the principal physical properties of platelets responsible for hemostasis and coagulation reactions. Chemically, they contain protein (60%), lipid (15%) and carbohydrate (8.5%). Their content of serotonin, epinephrine and norepinephrine aids in promoting constriction at the site of injury. The release of "platelet thromboplastin," a cephalin-type phosphatide, and ADP are important in blood coagulation.

As of the present time, there is no satisfactory manual method for accurate enumeration of blood platelets. The size and physical properties of the platelet seriously deter the development of accurate and reproducible methodology. Indirect methods of analysis are based on the proportion of platelets to erythrocytes in a stained blood smear. Blood samples obtained directly from the fingertip puncture are diluted with an anticoagulant fluid which simultaneously will stain the platelets. The ratio of platelets to red cells then is determined microscopically and the number calculated from the predetermined red-cell count (normal 3 to 8 platelets/100 RBC). In the direct procedures, a sample of blood is obtained by venipuncture, placed in a siliconized

tube, diluted and subsequently analyzed by counting the platelets in a microscopic counting chamber using conventional or phase-microscopy apparatus. Suitable diluting fluids are the Rees-Ecker Fluid (sodium citrate, 3.8 g; formaldehyde, 0.22 mL; brilliant cresyl blue, 0.05 g; water, qs 100 mL) or Brecker Fluid (1% ammonium oxalate). Automated procedures for platelet counting have increased the accuracy to ±5 to 10%. Blood is collected in a special anticoagulant, diluted and centrifuged at specified speeds to obtain a "platelet-rich" supernatant fluid, which then is counted in an automated counting apparatus similar to those used for RBC counting.

Methods for counting platelets in whole blood include electronic impedence instruments and laser-optical counters using hydrodynamic focusing.[3] These new hematology multiparameter analyzers provide greater accuracy, precision and increased rate of analysis performed on a small volume of blood. The automated instruments provide precise platelet measurements for monitoring chemotherapy-induced thrombocytopenia and transfusion therapy.

Persistent increases in platelet count (*thrombocythemia* or *piastrenemia*) have been observed in chronic myelocytic leukemia, polycythemia, megakaryocytic hyperplasia and splenic atrophy. Acute or temporary increases in platelet values (*thrombocytosis*) are seen in trauma and asphyxiation.

Thrombocytopenia or a decrease in platelets to values less than 60,000/cu mm occurs in various purpuras or hemorrhagic states (idiopathic or symptomatic thrombocytopenic purpura). Inherited platelet defects include Glanzmann's thrombasthenia which is characterized by prolonged bleeding time and poor clot retraction, while Bernard-Soulier Syndrome and Von Willebrand's disease demonstrates defective platelet adhesiveness. Defects in the release reaction includes "Storage Pool Deficiency" and "Aspirin-like" syndrome.

A rare, inherited, structural and functional platelet abnormality is the *grey-platelet syndrome* characterized by large platelets lacking alpha granules and appearing grey on Wright's-stained peripheral blood smears. Patients have a history of bleeding, petechiae, easy bruising and expistaxis. Diagnosis is confirmed by radioimmunoassay procedures to detect levels of platelet-specific alpha-granule proteins.

Leukemia, extensive burns, splenic disorders and agents such as quinidine, sulfonamides, hydrochlorothiazide, diuretics, antiepileptics and neuropharmacological agents have been implicated in the etiology of symptomatic thrombocytopenia. Decreases in platelet count also are accompanied by morphological changes in the size, shape and cytoplasmic granulation of these cells and changes in adhesiveness and normal function in hemostasis and coagulation.

Studies on *platelet aggregation* have been of significant value in the study of platelet abnormalities and their role in disease states. The rate and extent of the aggregation and clotting response to adrenaline, ADP, collagen and thrombin have been measured by observing changes in optical density of platelet-rich plasma on adding of these agents or other test substances. Low amounts of ADP give reversible aggregation, while a biphasic-aggregation pattern occurs with intermediate concentrations of ADP or with epinephrine. The second phase is the release of the platelets' endogenous ADP. High concentrations of ADP result in an irreversible aggregation. Aspirin acts as an inhibitor of the intrinsic-platelet ADP and the collagen reaction.

Reticulocytes—In normal peripheral blood 0.5 to 1.5% of the erythrocytes possess a fine reticulum in the cytoplasm. In blood smears prepared with Wright's, Giemsa and other Romanowsky methods, basophilic stippling of the erythrocytes occurs in lead poisoning (*plumbism*). This is not to be confused with the basophilic staining of the reticulocyte which only can be seen when cells are stained by supravital procedures (mixture of dyes with wet blood prior to preparing of an air-dried blood smear). The observed granular filaments or reticulum of this immature erythrocyte are a result of endoplasmic coagulation by lipophilic dyes used in the supravital procedures. *Reticulocytes* are enumerated by supravital staining of fresh blood with an anticoagulant-dye solution.

The usual method of expression is

$$\% \text{ Retics} = \frac{\text{No of reticulocytes/1000 RBC}}{10} \quad (1)$$

The "corrected" reticulocyte count is calculated for a more meaningful clinical approach in the degree of anemia by expressing the percentage of reticulocytes per mm^3 of whole blood.

$$\frac{\text{Corrected reticulocyte count}}{} = \text{Reticulocyte count} \times \frac{\text{(Patient's hematocrit)}}{\text{(Normal hematocrit)}}$$

In indirect counting methods a thin film of the blood-dye mixture is prepared on a microscope slide, counterstained with Wright's stain and the reticulocytes enumerated in proportion to a predetermined erythrocyte count. In direct procedures, reticulocytes are enumerated in wet films without counterstaining. Suitable dyes are brilliant cresyl blue, methylene blue and Janus green. These methods are subject to a high counting error.

An increase in the number of reticulocytes is an index of accelerated hematopoiesis and is observed in acute hemorrhage or adequate therapeutic management of iron-deficiency or pernicious anemia. In cases of chronic blood loss or bone-marrow depression a decrease in reticulocytes is seen.

Blood-Volume and Erythropoietic Mechanisms—The mean red-cell mass in normal males is 2095 ± 384 mL (30 mL/kg), the average plasma volume is 2766 ± 459 mL (40 mL/kg) and the total blood volume is 4861 ± 795 mL (70 mL/kg). The specific determination of *red-cell mass* is estimated accurately by tagging erythrocytes with ^{51}Cr *in vitro* or ^{59}Fe *in vivo*. These isotopes are incorporated into the β-polypeptide (Cr) or porphyrin (Fe) of hemoglobin in the RBC and subsequent isotope dilution in blood after injection of tagged erythrocytes is used for calculation of red-cell mass. In hemolytic anemia there is also a decrease in the normal life span (108 to 120 days) of the erythrocyte as indicated by a decreased survival time of ^{51}Cr-tagged red cells in blood (refer to Chapter 33).

Plasma volume is estimated by measurement of hemodilution of IV-injected ^{125}I or ^{131}I human serum albumin. The activity of labeled albumin steadily decreases after injection due to the loss of albumin to the extravascular space. Estimates of zero-time radioactivity levels can be made by extrapolation of a typical first-order blood-level decay curve. Dyes (Evans Blue) and other isotopes are less satisfactory for accurate assessment of plasma volume. The total blood volume is equal to the red-cell mass and plasma volume.

Chronic expansion of the red-cell mass is seen in primary and secondary polycythemia associated with erythrocytosis due to hypoxia, tumors and renal disease. In these conditions, there is an increased hemoglobin and hematocrit and absolute increase in red-cell mass. In relative polycythemia the high hematocrit is due to contraction of the plasma volume. *Chronic expansion of the blood volume*, with a resultant decrease in hematocrit value and, in some cases, a "hemodilution" anemia, is seen in cardiac failure, normal pregnancy, hepatic cirrhosis, splenomegaly and arteriovenous fistula.

The metabolic defect in *pernicious anemia*, characterized by inadequate gastrointestinal absorption of vitamin B_{12}, is diagnosed readily by monitoring urinary radioactivity fol-

lowing oral administration of cyanocobalamin-^{57}Co with and without intrinsic factor. The percent recovery of the isotope in normal patients is 3 to 25% and in pernicious anemia 0 to 2.5%.

^{51}Cr-tagged erythrocytes also are used in studying the effects of various compounds, such as the nonsteroidal anti-inflammatory drugs, on *gastrointestinal (GI) bleeding*. The patient's blood cells are tagged with ^{51}Cr and the agent under test is administered. If GI bleeding occurs, there is an increase in ^{51}Cr content of fecal samples as a result of blood loss into the lumen of the GI tract.

Measurement of the absorption of radioactive iron (^{59}Fe), its tissue distribution (liver, spleen, precordium, sacral bone marrow), plasma elimination and urinary excretion establish various *ferrokinetic parameters*. Iron is absorbed to the greatest extent as the ferrous salt in the upper small intestine. Absorption is decreased in iron overload, erythropoiesis and various malignant, inflammatory or infectious diseases. Iron is transported in plasma bound to transferrin, a specific iron-binding protein. Alterations in plasma iron and iron-binding capacity are seen in pregnancy, thalassemia major and iron deficiency (hypochromic) anemia. Iron is stored in the liver, bone marrow, skeletal muscle and spleen as ferritin and hemosiderin. The daily turnover of iron is about 35 mg, primarily from an "erythropoietic labile pool" in bone marrow.

Hemosiderosis is simply an increase in iron storage, whereas *hemochromatosis* denotes increased iron storage with associated tissue damage. Both of these states can result from oral or parenteral medicinal/transfusion iron overload. Iron excretion is limited and occurs by desquamation of iron-containing cells from the bowel, skin and urinary tract.

Iron-deficiency anemia is a symptom and not a disease. Treatment is based on evaluation of ferrokinetic parameters, correction of hemoglobin and tissue-iron deficiency and recognition of the underlying cause (eg, chronic blood loss).

Blood Coagulation—*Hemostasis*, the arrest of blood flow from a vessel, is regulated by extravascular (muscle, skin and subcutaneous tissue), vascular (blood vessels) and intravascular (platelet-adhesion, clot-retraction and blood-coagulation) mechanisms. The following discussion will be limited to those processes related to the blood-coagulation mechanism. When blood is allowed to clot, the free-flowing liquid is converted into a firm cell clot surrounded by serum. If an anticoagulant is added to blood, coagulation does not occur and the blood cells are suspended in a liquid phase —plasma. The clotting mechanism involves three stages: the formation of plasma *thromboplastin*, the conversion of *prothrombin* to *thrombin* and the conversion of *fibrinogen* to *fibrin*.

The International Committee on Nomenclature of Blood Clotting Factors has numerically designated the blood-coagulation factors (Table II). Fibrinogen and Factors V and VIII are absent in normal blood serum as a result of the clotting process. The absorption characteristics of certain blood-coagulation factors on calcium phosphate or barium sulfate are used in the differential analysis of specific factors. The interaction of coagulation factors may be initiated through either the intrinsic or extrinsic pathways. In the intrinsic system all the factors are present in the blood, while the extrinsic system is activated by the release of tissue thromboplastin. Figure 28-3 shows the activities of both pathways to form a stabilized fibrin clot.

In Stage 1 of the coagulation process, the contact of injured tissue with blood results in the activation of Factor XII, which reacts with calcium, PTA, PTC, AHG and Factors III, V and X to yield intrinsic or blood thromboplastin. This stage normally is completed in 3 to 5 min. Extrinsic or tissue thromboplastin is formed rapidly (<12 sec) in various tissues in the body such as lung and brain in the presence of calcium and Factors V, VII and X.

Table II—Blood-Coagulation Factors

Factor	Synonym
I	Fibrinogen
II	Prothrombin
III	Thromboplastin (tissue)
IV	Calcium
V	Labile factor, proaccelerin, Ac globulin
VI	Accelerin
VII	Stable factor, proconvertin, serum prothrombin conversion accelerator (SPCA)
VIII	Antihemophilic globulin (AHG)
IX	Christmas factor, plasma thromboplastin component (PTC)
X	Stuart-Prower factor
XI	Plasma thromboplastin antecedent (PTA)
XII	Hageman factor
XIII	Fibrin-stabilizing factor (FSF)

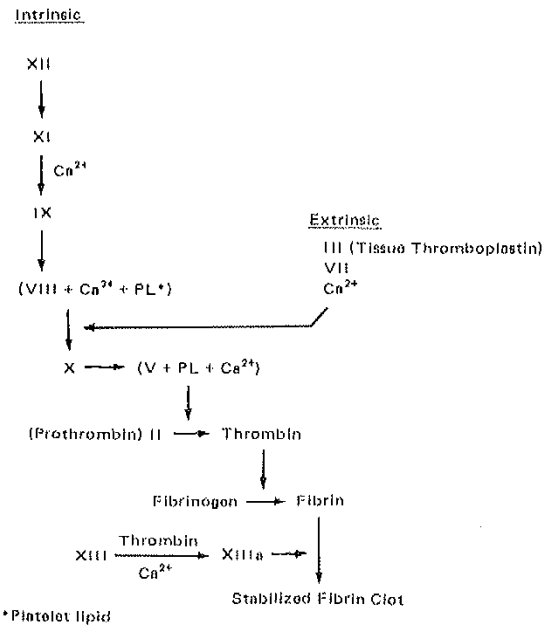


Fig 28-3. Blood Coagulation Process.

In Stage 2, thromboplastin catalyzes the conversion of prothrombin to thrombin (8 to 15 sec) in the presence of Factors V, VII, X and calcium.

In Stage 3, the thrombin rapidly converts fibrinogen into fibrin, which then forms a network of fibers that traps red cells and thus forms the blood clot.

Although the exact nature of the enzymatic sequences in the coagulation process is not clear, it is definitely a biological amplification process starting from the small reaction of tissue contact to rapid conversion of fibrinogen to fibrin.

Blood contains natural inhibitors of coagulation such as antithrombin, heparin and antithromboplastin, which can prevent a particular reaction in the coagulation sequence. The dissolution of blood clots occurs by the action of the blood proteolytic enzyme—plasmin or fibrinolysin. Plasmin is formed from its precursor, plasminogen, after activation by tissue and body fluids or substances of bacterial origin (streptokinase).

The routine tests performed in the coagulation laboratory are indices of vascular function (vascular phase and platelet adhesion) or intrinsic clotting mechanisms. Determinations of *bleeding time* and *capillary fragility* provide esti-

mates of blood coagulation in the presence of platelets and tissue or vascular factors. In the Ivy method for determination of *capillary bleeding time*, a blood pressure cuff is placed on the forearm and inflated to 40 torr; a puncture wound is made and the time required for bleeding to stop is noted. *Bleeding time* is a screening test for disorders of platelet function or vascular defects but is usually normal in coagulation disorders. The test is useful in the differential diagnosis of Von Willebrand's (reduced factor VIII, with a normal bleeding time) disease from mild hemophilia. The normal bleeding time, as determined by this method is 1 to 9 min. Dextran, pantothenyl alcohol, and derivatives, penicillin G, nonsteroidal anti-inflammatory drugs and streptokinase-streptodornase may cause a prolonged bleeding time. The *Simplate II* (General Diagnostics Div, *Warner Lambert*) is a standardized, disposable, springloaded bleeding-time device for platelet function testing. It uses two blades that are released automatically to produce two uniform incisions 6 mm long × 1 mm deep, making the procedure reliable and reproducible.

The *capillary fragility* or *tourniquet test* is based on the incidence of petechiae (small red marks) formation produced by an inflated blood pressure cuff over a 5-min period. Normally, a few tiny petechiae may appear. The most common cause of abnormalities in vascular-function and platelet-adhesion tests is thrombocytopenia.

An analysis of the *intrinsic coagulation mechanism* is concerned with the determination of the levels of the specific clotting factors in whole blood. In preliminary studies of a suspected hemorrhagic disorder, determinations of *coagulation time, clot retraction, platelet count, bleeding time* and *capillary fragility* usually are performed.

In the Lee-White procedure, the coagulation time of whole blood is determined in regular or siliconed tubes. Normal values are 8.5 to 15 min in glass and 19 to 60 min in siliconed tubes. Anticoagulants and tetracyclines may cause increased times while corticosteroids and epinephrine cause decreased values. The siliconization of glassware prevents platelet aggregation and thus, delays coagulation. The samples used in the analysis of coagulation time are then inspected at 0.5, 1, 2, 4 and 24 hr after clotting to determine the time required for the various phases of clot retraction. The tubes also are observed for evidence of clot lysis or dissolution. The clot normally will start to retract in 30 min, completely retract within 24 hr and show no evidence of lysis over a 72-hr period. Prolonged coagulation times are associated with hemophilia, hypofibrinogenemia and Factor IX deficiency. Abnormalities in any of these tests indicate the requirements for further coagulation studies.

The *prothrombin time test* is a measure of the levels of all coagulation factors, except III, IV and VII, and is an index of the capacity of plasma to form thrombin. In the "One Stage" test, the plasma sample is mixed with calcium chloride and tissue thromboplastin, and the time required for fibrin-clot formation is determined. Results are compared with a normal plasma control, and the prothrombin time is reported either in seconds or as the percent of prothrombin calculated from a standard activity curve. Correction studies using normal serum, adsorbed normal plasma or whole normal plasma added to test serum indicate deficiencies of Factors VII and X, Factor V and Factor II, respectively. If none of these additives shorten the prothrombin time, a circulating anticoagulant problem can be suspected.

A modification of this technique (the *prothrombin-proconvertin procedure*) using a 1:10 dilution of both patient and control plasma in the presence of prothrombin-free plasma as a source of Factors I and V, is a more sensitive index of specific deficiencies in prothrombin, Factor VII, IX and X.

Owren's *thrombotest*, as performed on whole blood, is sensitive to changes in both extravascular and intravascular clotting mechanisms, including Factor IX. The dosage of anticoagulant drugs, such as dicumarol, is adjusted in accordance with prothrombin-time determinations; patients are maintained usually within a therapeutic range of 20 to 40% of prothrombin activity (normal range, 80 to 130%). Reduced prothrombin levels, with prolonged prothrombin times, are observed in vitamin K deficiency, hemorrhagic disease of the newborn, excessive anticoagulant therapy, liver and biliary disease. The interaction of other drugs with anticoagulants may cause increased prothrombin times. Drugs such as salicylates, phenylbutazone, oxyphenbutazone, indomethacin and some sulfonamides increase the amount of active anticoagulant activity. Other drugs decrease the amount of vitamin K produced by gut bacteria which include chloramphenicol, kanamycin, neomycin, streptomycin and the sulfonamides.

The *prothrombin consumption test* is an index of the efficiency of conversion of prothrombin to thrombin in the coagulation process. The blood sample is allowed to clot under standardized conditions and then the quantity of prothrombin complex removed in the serum is determined in the presence of extrinsic fibrinogen. At least 80% of the prothrombin is consumed normally. Reduced consumption of prothrombin (<80%) is observed in coagulation deficiencies (hemophilia) related to thromboplastin generation.

Other types of coagulation tests detect deficiencies in *thromboplastin generation mechanism*. The *thromboplastin generation time test* (TGT) provides a means of detecting specific deficiencies of Factors V, VIII, IX, X, XI or XII. In the initial phase of this procedure the clotting time of the patient's adsorbed plasma is determined in the presence of a standardized platelet factor reagent, calcium chloride, plasma substrate reagent (Factors I, II and V) and the patient's serum. If the clotting time is abnormal (>16 sec), further tests are performed with the patient's plasma or serum. The adsorption of the plasma sample on barium sulfate removes Factors II, VII, IX and X and facilitates differentiation of a Factor IX to X from V to VIII deficiency in the thromboplastin-generation mechanism. Thromboplastin generation is reduced in hemophilia and thrombocytopenia.

The *activated partial thromboplastin time test* (PTT) is based on the observation that hemophilic plasma has a normal clotting time in the presence of a complete thromboplastin (extrinsic-saline extract of brain tissue), as used in prothrombin determinations, but will give a markedly prolonged clotting time with an incomplete thromboplastin (cephalin). Cephalin is a thromboplastic, ether-soluble phospholipid factor with platelet-like activity. In this test the clotting time of the patient's plasma is determined in the presence of calcium chloride and activated cephalin. This test is used primarily to detect deficiencies in Stage 1 of the coagulation mechanism and is rather sensitive to changes in Factors VIII and IX, as seen in classical hemophilia and Factor IX deficiency (Hemophilia B or Christmas disease).

In Stage 3 of the coagulation process, the presence of adequate levels of fibrinogen and thrombin is critical. *Fibrinogen levels* are analyzed semiquantitatively by determining the clotting time of a diluted plasma sample in the presence of extrinsic thromboplastin. This test is basically independent of prothrombin levels. Fibrinogen concentrations of 125 mg% or greater are adequate; deficiencies (hypofibrinogenemia) have been observed in liver disease, carcinomatosis and in certain complications of pregnancy.

Increased levels of *fibrinogen degradation products* (FDP) have been demonstrated in serum due to primary activation of the fibrinolytic system (pathological fibrinolysis) or by secondary activation following increased blood clotting (disseminated intravascular coagulation). Fibrinogen (mol wt 3.4×10^6) is degraded sequentially to fragments

X, Y, D and E with mol wts of 2.7, 1.65, 0.85 and 0.55 × 10^5, respectively. Fragments X and Y are more potent anticoagulants than fragments D and E and are responsible for hemorrhagic states in defibrination. Complexes between fibrin monomer, fragment X and other FDP interfere with thromboplastin generation and platelet formation. FDP can be measured by immunological techniques involving latex agglutination of particles sensitized with specific antibodies to FDP or by a hemaglutination-inhibition test. The normal level of serum FDP is 4.9 ± 2.8 μg/mL. Increased levels are seen in acute myocardial infarction, menstruation, complications of pregnancy, hypoxic newborns, malignancy and renal disease.

Deficiencies in the clotting mechanisms usually can be corrected partially and temporarily by transfusion of normal blood or plasma. When this fails, the presence of *circulating anticoagulants* (antithrombin, antithromboplastins, heparin) must be considered. Heparin acts indirectly by means of antithrombin III, which neutralizes several activated clotting factors (XIIa, activated Fletcher factor, XIa, IXa, Xa, IIa and XIIIa). The pharmacological effect of an oral anticoagulant is the inhibition of blood clotting by interfering with vitamin K-dependent clotting factors II, VII, IX and X. Circulating anticoagulants are detected by determining the effect of normal plasma on the clotting time (*recalcification time*) of the patient's oxalated plasma in the presence of calcium chloride. If the addition of the normal plasma does not shorten the prolonged recalcification time, a circulating anticoagulant state can be reported.

Since the end-point of all coagulation tests is the conversion of fibrinogen to fibrin, it is vital that the analyst rigidly standardize his concepts of fibrin formation in visual recording procedures. The use of mechanical instrumentation in the detection of clot formation significantly has increased the standardization, accuracy and reproducibility of coagulation procedures. These instruments measure and record the process of fibrin formation via increased turbidity (coagulogram or photometric clot detection) or changes in electrical conductance in the reaction mixtures. As well as performing routine laboratory tests simultaneously or sequentially, updated systems can run Fibrinogen and Factor assays achieving rapid throughput and accuracy. New performance features are available with many of the automated coagulation instruments. These include monitoring temperature zones, digital displays of the individual clotting times, automatic dilutions or patients samples and programmable parameters for testing flexibility.

Hemophilia is a classic deficiency of antihemophilic globulin (AHG), Christmas disease of PTC and Hageman trait of Factor XII. Hereditary or acquired deficiencies of Factors II, V, VII, X and XI also are associated with disease states. The process of blood coagulation, analysis of coagulation factors and interpretation of results comprise a highly complex system. The coagulation laboratory and the physician function together in the diagnosis and treatment of coagulation-deficiency diseases.

Blood-Bank Technology

Blood-bank technology in the modern laboratory is part of the blood-transfusion service. As whole blood for transfusion and its components are biologically active therapeutic substances, a complete analysis of their chemical and biological characteristics is vital to the assurance of successful therapeutic effects. The transfusion service is responsible for:

1. Receiving and examining of the donor.
2. Collecting, processing and storing the blood.
3. Typing of recipient and donor for ABO and Rh blood-group factors.
4. Compatibility (cross-matching) testing before transfusion.
5. Issuing of blood for transfusion and extracorporeal circulation.
6. Evaluating transfusion complications.
7. Performing of special serological tests pertinent to blood groups and other factors.

In this section a discussion of pertinent factors related to the various phases of the transfusion service will be presented.

Receiving and Examining of the Donor—A complete registry[4] of prospective donors should be maintained, with specific reference to age, sex, weight, address, occupation and telephone number. Computerized blood banking has increased the efficiency of this service. Donors should preferably be between the ages of 21 and 60 and should weigh no less than 110 lb. The donor may be rejected on the basis of previous or active incidence of certain microbial diseases (recurrent malaria, syphilis, infectious or homologous serum hepatitis, tuberculosis), bleeding abnormalities, convulsions, allergic syndromes, skin or heart diseases, diabetes, alcohol or drug addiction, pregnancy, cancer, recent immunization with live vaccine product, acquired immune deficiency syndrome (AIDS) or blood-pressure abnormalities (acceptable blood pressure: between 100/50 and 200/100; pulse rate: 60 to 12/min). The screening of blood for exposure to human immunodeficiency virus (HIV) is crucial to reducing the risk of infection from transfusion. ELISA (enzyme-linked immunosorbent assay) screening tests for the detection of antibodies against HIV are available from manufacturers. More sensitive tests are being developed to detect viral DNA in body fluids.

A period of at least 8 weeks should have elapsed since blood was withdrawn and the blood hemoglobin level should be 12.5 to 13.5 g% or greater. Serum bilirubin and transaminase levels also should be evaluated in donors with previous incidence of jaundice.

Collecting, Processing and Storing the Blood—A tourniquet is applied to the arm of the donor to occlude the venous return, the skin area is sterilized and the blood is collected by venipuncture (phlebotomy). NIH Formula A or B [ACD(Acid–Citrate-Dextrose) or ACD-phosphate] solutions are used as anticoagulants in the sterile blood-collecting containers. Evacuated containers may be of regular or siliconed glass; collapsible plastic containers offer many advantages in donation, blood-banking and transfusion procedures.

The preservation of the red cells in blood is improved by the complete removal of trapped air in the blood-collection apparatus, rapid cooling after collection and storage at 4°. Properly collected whole blood is usually stable for 21 days at 1 to 6°. The deterioration of whole blood is related to increased cellular fragility (increased plasma K^+) and decreased glucose utilization. Blood which is used for correction of any bleeding tendency or clotting defect should be as fresh as possible. Leukocytes, platelets and Factors V and VIII deteriorate in stored plasma or whole blood.

ABO Blood-Group Classification[5]—Human red cells can be classified into various groups or types on the basis of reactivity of certain blood factors (*agglutinogens*) located on the erythrocyte membrane. The Landsteiner system (Table III) for the four blood groups is based on the presence or absence of either A or B agglutinogen on the cell surface (Group A, B, AB or O, respectively).

Serum does not contain the antibody (*agglutinin*-IgM type) for the antigen present in an individual's own red cells, but does contain the isoagglutinin (eg, anti-B in blood group A) due to exposure, early in life, to bacterial and plant antigens similar in structure to the A-B antigens. The clumping or agglutination of the red cells by reaction of agglutinogen with agglutinin is used in blood-grouping techniques. In certain instances hemolysin antibodies, present

Table III—Blood-Group Systems

Blood Group	Agglutinogen in Cell	Agglutinin in Serum	Reaction[a] with Anti-A Serum	Reaction[a] with Anti-B Serum	Frequency (%) in Caucasians
A	A	Anti-B	+	–	41
B	B	Anti-A-A_1	–	+	10
AB	AB	None	+	+	4
O	None	Anti-A and B	–	–	45

[a] Agglutination.

in serum containing anti-A or anti-B agglutinins, cause the disruption of cells and release of hemoglobin (hemolysis).

Human blood cells are grouped by two separate reactions: cellular or "front" grouping and serum or "reverse" grouping. The blood group ordinarily is determined by testing an individual's red cells with standardized anti-A or anti-B serum (certified by the Div of Biological Standards, NIH). Confirmation of the blood group (reverse typing) is accomplished by an analysis of an individual's agglutinin titer. In this procedure the individual's serum is heated at 56° for 10 min to destroy hemolysins, and then mixed with known Subgroup A_1 or B_1 human red (Rh-negative) cells in the agglutination test. These two tests should be in agreement prior to the release of blood for transfusion.

Although human blood cells of Group B react uniformly with Anti-B serum, Group A and AB cells show a wide range of reactivity with Anti-A or Anti-A_1B serum. Blood-group A may be further categorized into Subgroups A_1, $A_{int.}$, A_2, A_3, A_0 and A_x on the basis of the reaction with absorbed Anti-A, Anti-A_1-lectin, Anti-H-lectin, Anti-$A_{1,2}$ and Anti-AB serum and the presence of Anti-A_1 in the serum. Certain Group O individuals possess anti-H in their serum and are further subcategorized into the Bombay or O_h phenotype. Tests for A, B and H in saliva can establish the genotype of an individual, ie, A and H in saliva of blood-group A, B and H in B, H and O and A, B, H in AB. This is helpful in cases of poorly developed red-cell antigens or in the loss of cellular antigen in some patients with leukemia.

As the human blood cell contains many antigens with rather complex biochemical and immunochemical properties, the blood factors have been classified further into various subsystems. The Kell (K), Lutheran (Lu), Lewis (Le), Duffy (Fy), Kidd (Jk), MNS, Sutter (Js), Diego (Di) and P blood-factor systems are based on the detection of a specific antigen on or within the red cell by means of antibody (*isohemagglutinin*) reactions with specific antisera or panels of reagent red cells. Some of these factors (eg, Kidd, Kell and Lewis) have been involved in transfusion reactions.

The Rh-Hr System and Antihuman Globulin Test—The presence or absence of *Rh_0 antigen* in human blood is of prime importance in transfusion reactions, paternity disputes and isosensitization phenomena. There are eight blood Rh phenotypes which are determined by their reaction with three specific serum agglutinins (Anti-Rh_0, Anti-rh′ and Anti-rh″): rh, rh′, rh″, rh′rh″, Rh_0, Rh_0′, Rh_0″ and Rh_0′Rh_0″. The rh groups do not contain the Rh_0 factor on the cell surface and are designated "Rh-negative." The terminology of the Wiener system (Rh, rh) is comparable to the Fisher-Race (CDE) as follows: rh′(C), Rh_0(D), rh″(E). The Rosenfeld system uses a numerical classification: RH1 = Rh_0.

The absence of the Rh antigen in about 15% of the population does not preclude the presence of other factors; the use of specific antisera (Anti-hr′ and Anti-hr″) has demonstrated the existence of the Hr factors (Hr_0, hr′, hr″). For example, the Rh-negative cell (rh″) possesses rh″hr′Hr_0 antigens. The antigen Rh_0(D) is the most potent immunogen of all the Rh antigens.

The Rh antibodies are either *saline agglutinins* (complete) or "*blocking*" *antibodies* (incomplete). The latter are of the IgG type. They are used in Rh testing procedures and are produced more commonly, and in higher titer, in the human isosensitization or autoantibody reactions. They will not agglutinate saline suspensions of normal Rh-positive red cells except in the presence of a high concentration of albumin, serum or conglutinin (AB serum with albumin) at a temperature of 35 to 37°.

In routine Rh testing procedures, a sample of blood (oxalated or heparinized) or a suspension of cells in serum or albumin is mixed with Anti-Rh_0 serum on a slide or in a tube at 37 to 47°. The presence of clumping indicates that the blood possesses Rh_0 antigen. Confirmation of an Rh-negative test may be performed by retesting with Anti-rh′Rh_0rh″ serum.

In Rh testing procedures, red cells from patients with acquired hemolytic anemia are partially coated with human autoantibody, and cells from erythroblastic infants are coated with maternal antibody globulins and may be clumped falsely by Rh typing serum containing a high protein concentration, or may appear to be Rh-positive in the saline-cell suspension test. Demonstration of anti-Rh_0(D) in an eluate from these antibody-coated cells can help to establish true Rh type.

Anti-Rh antibodies are not normally present in human serum; they may be acquired via isosensitization. The transfusion of Rh-positive blood to an Rh-negative recipient, or transfer of cells of Rh-positive fetus through the placental barrier to the Rh-negative mother, will result in formation of antibodies to Rh agglutinogens not present in the cells of the recipient or mother, respectively.

Hemolytic blood-transfusion reactions and hemolytic disease of the newborn (erythroblastosis fetalis) involve *isosensitization phenomena* usually related to the Rh_0 antigen. Hr and ABO antigens also can be responsible for hemolytic disease of the newborn. If an expectant mother is Rh-negative and the father is Rh-positive, the Rh genotype of the father should be determined. If the father is homozygous, the erythrocytes will contain a pair of Rh_0 factors and the offspring will inherit the Rh_0 factor; if he is heterozygous, one Rh_0 and one Hr_0 factor will be present and his offspring may or may not inherit the factor.

If the fetus is Rh-positive, the mother may be sensitized to the Rh antigen and in subsequent pregnancies the development of high titers of Anti-Rh_0 antibodies will result in hemolytic disease of the fetus. These antibodies enter the fetal circulation via the placental barrier, coat the red cells of the fetus and cause excessive erythrocyte destruction, hyperbilirubinemia and associated potential for brain damage, hydrops fetalis (edema) and congenital anemia of the newborn. This Rh disease can be avoided now by proper therapeutic use of Rh_0(D) Human Immune Globulin (RhoGAM, *Ortho*) to prevent the postpartum formation of active antibodies in the Rh_0(D)-negative, D^u-negative mother who has delivered an Rh_0(D)-positive or D^u-positive infant.

The *Coombs' antiglobulin test* is a method of detecting the blocking-type antibodies, globulins and complement which are attached to red-cell antigens in isosensitization phenomena.

In the "direct" test procedure, a saline suspension of washed red cells is mixed with anti-human gamma globulin antiserum and agglutination is indicative of the combination of human antibody with antigen on the red cell, eg, maternal incomplete isoantibody on infant's red cells in hemolytic disease of the newborn, autoimmune, drug-induced, alloantibody-induced hemolytic anemia and after transfusion of incompatible red cells.

An "indirect" procedure is used to demonstrate the presence of blocking antibody in the serum of pregnant Rh-

negative women and in transfusion reactions. In this procedure the patient's serum is incubated with a suspension of Group O Rh-positive red cells; the cells are washed and then antihuman globulin antiserum is added to detect the coating of the red cells with antibody globulin from the patient's serum by agglutination phenomena. If agglutination occurs in the first part of the procedure, a saline agglutinin is also present. Anticomplement sera (anti-nongammaglobulin antiserum) are used to detect reactions involving anti-JK.

The Du allele is a clinically important variant of the Rh_0 factor and usually associated with rh′(C) and rh″(E). Individuals with this factor are considered Rh-positive, and the red cells fail to react with anti-Rh_0 in the saline-tube method but reacts with incomplete anti-Rh_0(D) by other slide or tube techniques. Rh-negative donors should be tested for Du factor. If positive, their blood must only be given to Rh-positive recipients.

Drug-Related Problems—Hematological abnormalities may be caused by the administration of drugs which can cause a positive direct antiglobulin test and immune hemolytic anemia, eg, cephaloridine, cephalothin (*Keflin*), methyldopa (*Aldomet*), penicillin, L-dopa, quinidine, phenacetin and insulin.

Compatibility Testing—Cross-matching procedures are designed to detect incompatibilities in the blood of donors and recipient. The test is designed to prevent transfusion reaction and assure maximum benefit to the patient. Although erroneous ABO grouping usually will result in an incompatible cross match, no such protection exists in the Rh system. An incorrectly typed Rh-positive donor blood can result in primary immunization to Rh_0(D) antigen if transfused to an Rh-negative recipient. For each transfusion, a *major* and *minor cross match* should be performed.

In the *major cross match* (1) a saline suspension of the donor's cells is mixed with the recipient's serum and (2) the donor's cells are suspended in recipient's serum or in serum with added albumin. The saline cross match is an additional check on the ABO typing and may detect incompatibilities caused by antibodies to M, N, S, P and Lu subgroups. The high-protein or albumin cross match can demonstrate antibodies in the Rh system. The presence of agglutination or hemolysis indicates incompatibility.

The *minor* cross match includes the donor's serum and the recipient's cells, and is useful as a check of the ABO typing and an indication of the possibility of transfusion reactions caused by a rare antigen on the recipient's cells or uncommon antibodies directed against an antigen in the serum of the donor. The minor cross match has been replaced in many instances with screening of the donor's serum against a panel or pool of red cells of known antigenicity.

The *indirect antihuman globulin* procedure also must be performed with the recipient's serum and donor's cells with and without albumin (major side) and may be tested with the donor's serum and recipient's cells (minor side). The use of proteolytic enzymes (bromelain) enhances the agglutination of red cells by low-titer or weakly reacting Rh-Hr antibodies, probably by removing sialic acid residues on the RBC surface. The red cells used in the indirect Coombs test are treated with the enzyme prior to absorption of antibodies and addition of antiglobulin reagent.

The usual cross-matching techniques involve (1) a room-temperature or 30° procedure, preferably with the addition of albumin, (2) a high-protein procedure and (3) an antiglobulin procedure.

The presence of nonspecific *autoantibodies*, *cold agglutinins* and *bacteriogenic agglutination* sometimes complicates the cross-matching procedure. If the recipient's serum reacts more strongly with his own cells than with the donor's, autoantibodies should be suspected. Cold agglutinins usually will agglutinate all blood, regardless of type, at low temperatures, but will not react at 37°. Agglutination as a result of bacterial contamination of blood is called panagglutination.

Hepatitis Testing—Posttransfusion hepatitis is associated with the transmission of virus-like particles referred to as *Australia or serum hepatitis antigen or the hepatitis associated antigen (HAA)*. All donor blood must be tested for the presence of HAA. Agar gel diffusion (AGD), counterelectrophoresis (CEP), complement fixation (CF) and rheophoresis procedures can be used.[6] The rheophoresis procedure uses a modified gel-diffusion technique for the detection of HAA by precipitin-type reaction with HAA antibody. It offers the sensitivity of CEP and CF procedures with the simplicity of the AGD procedure. Other tests for HAA are based on radioimmunoassay (RIA) technique for detection of antigen by hemagglutination (HA) or HA-inhibition for the presence of HAA antibody. In the RIA technique, the donor's serum is added to a test tube coated with HAA antibody (solid RIA). If the serum contains HAA, it will bind to the antibody. ^{125}I-HAA is then added to the tube. If the antibody binding site is occupied previously with HAA from the donor's serum, ^{125}I-HAA will not bind and the determination of ^{125}I bound versus free is an index of HAA content of the donor's serum.

Issuing of Blood and Evaluating Transfusion Reactions—Whole-blood, red-cell or leukocyte suspensions, plasma, platelet-rich plasma, platelet concentrates, leukocyte-poor blood, AHF, factor IX complex, plasma protein fractions and RhoGAM are products of the transfusion service.[7] Transfusion reactions are related to antibody phenomena or disease transmission. The hemolytic reaction resulting from the transfusion of incompatible cells is the most serious problem. The transfusion of microbially contaminated blood can result in a pyrogenic reaction or transmission of infectious diseases, such as malaria, syphilis or hepatitis. Allergic reactions (urticaria, asthmatic seizures), circulatory overload, embolic complications (blood clot, air emboli) also may be encountered. Leukocyte and platelet antibodies develop in repeat transfusions and in transplantation patients. The transfusion service is an integral unit in evaluating such complications.

Techniques of Analysis

This section will describe the principles of the procedures used in the analyses of various substances in blood, plasma or urine. Examples of the significance of such tests in clinical diagnosis will be presented. For a complete description of the physiological and pharmacological aspects of these blood constituents, see the *Bibliography*.

Instrumentation—The development of instrumentation has accelerated progress in clinical chemistry. An excellent review of the principles and applications in clinical chemistry of automation, atomic-absorption spectroscopy, ultraviolet and visible spectrophotometry, fluorimetry, phosphorimetry, infrared and Raman spectroscopy, microwave and radiowave spectroscopy and nucleonics was prepared by Broughton and Dawson.[8] Quality-control techniques are a vital part of any clinical laboratory. Standard reference materials,[9,10] standardization of quantities and units[11] and continual evaluation of precision and accuracy of various determinations[12] are incorporated into procedures of all reliable clinical laboratories. The manufacture of certified standards and reagents and the certification of clinical chemists and clinical laboratories are under the supervision of either the FDA, NIH, Pharmaceutical Manufacturers Association (PMA), American Association of Clinical Chemists, the College of American Pathologists and the National Committee for Clinical Laboratory Standards (NCCLS).

Interaction of Drugs with Clinical Laboratory Tests—Drugs may interfere with the interpretation of laboratory tests by three classes of mechanisms:

I. *Chemical or biochemical* interference due to reaction of a drug or its metabolite in biological fluids with test reagents in analytical procedures.
II. *Pharmacological* interference due to normal drug-induced alterations in various physiological parameters.
III. *Toxicological* interference as a consequence of the toxicity of a drug.

Examples of Class I interference include false-positive urine glucose results due to the reducing properties of drugs or metabolites such as ascorbic acid, *p*-aminosalicylic acid, tetracycline, cephaloridine and levodopa, which are excreted in urine. Spironolactone will result in an elevation of certain urinary ketosteroids through cross-reaction of the drug in the analytical procedure.

Examples of Class II interference include the decrease in serum-potassium levels in patients receiving thiazide diuretics, the alteration in serum uric acid with probenecid and the elevation in various plasma proteins and thyroid function tests with estrogen-progesterone combinations. Drug-drug interaction also can result in changes in these parameters. Guanethidine enhances the effect of the coumarin anticoagulants. Barbiturates induce hepatic microsomal enzyme synthesis and subsequently increase the metabolism and decrease the therapeutic effect of drugs, such as warfarin, even after these drugs are terminated.

Examples of Class III interference include changes in liver- and kidney-function tests and hematological parameters (anemia, agranulocytosis, leukopenia) due to drug-induced toxicity and positive LE and ANA tests due to a "lupus-like" syndrome induced by hydralazine.

It is beyond the scope of this chapter to include a complete listing of drug interactions in laboratory tests. The reader is referred to an annual, readily available, computerized review of the effect of normal therapeutic drug doses, as well as overdoses, on clinical laboratory tests[13] and to other review articles.[14]

Blood

Collection and Preparation for Chemical Analysis—Using aseptic technique, a blood sample is obtained by venipuncture and usually placed in evacuated glass tubes. The choice of anticoagulant, type of specimen, stability of test component and use of preservatives depends on the type of analysis requested and the specific analytical procedure involved. If serum is desired, the blood sample is allowed to clot and the serum is separated by centrifugation. When whole blood or plasma is to be used in the analysis, an anticoagulant is added to the collecting tube.

The following concentrations of specific anticoagulants are used routinely per 10 mL blood; lithium, potassium or sodium oxalate (15 to 25 mg), sodium citrate (40 to 60 mg), heparin sodium (2 mg), disodium or tripotassium ethylenediaminetetraacetate (EDTA-Na_2, 10 to 30 mg) or ACD-Formula B solution (1.0 mL).

Heparin prevents blood coagulation by inhibiting the thrombin-catalyzed conversion of fibrinogen to fibrin. The other anticoagulants either precipitate blood calcium or convert ionized calcium into a nonionized (chelated) form which cannot function in the coagulation reaction. Heparin and EDTA do not alter the cellular elements of blood significantly. Sodium fluoride and thymol are used as preservatives or enzyme inhibitors to prevent the deterioration of various substances in the blood sample, eg, glucose → lactic acid. Preservatives and anticoagulants can interfere with some enzyme tests. Serum usually is used for these procedures.

The separation of plasma or serum, and chemical analysis, usually are performed as soon as possible after the collection of the sample. The addition of polystyrene granules to the blood sample prior to centrifugation facilitates the isolation of serum or plasma. Hemolysis interferes with analytical procedures for bilirubin, albumin, nonprotein nitrogens, pH, phosphorus, potassium and various enzymes. The serum also should be observed for presence of lipemia. Changes in the ratio of CO_2, chloride and electrolytes in cells and plasma, glycolytic conversion of glucose to lactic acid, hydrolysis of ester phosphate to free inorganic phosphate, bacterial conversion of urea to ammonia and conversion of pyruvate to lactate are examples of changes that can occur in contaminated, improperly preserved or unrefrigerated blood specimens.

The first stage in many of the chemical determinations is the removal of blood protein and preparation of *protein-free blood filtrate*. The protein is precipitated with tungstic acid, trichloroacetic acid, zinc hydroxide or organic solvents, such as alcohol and acetone, and then filtered or centrifuged to remove the protein coagulum. Tungstic acid precipitation is performed by mixing 1 volume of blood or 2 volumes of plasma with 9 volumes of stabilized tungstic acid reagent. The filtrate obtained in this procedure should be in the pH range of 3.0 to 5.1 to assure the adequate removal of proteins (<2 mg% in filtrate).

The Somogyi filtrate is prepared by mixing 1 volume of blood with 5 volumes of water, 2 volumes of 5% zinc sulfate and 2 volumes of 0.3 *N* barium hydroxide. The barium sulfate is precipitated and the zinc hydroxide formed in the reaction precipitates the blood proteins. Trichloroacetic acid (10%), in a ratio of 9:1 with blood, yields greater volumes of filtrate due to a more complete formation of protein agglomerates.

Blood Glucose—Methods for determining blood glucose are based on the use of glucose as a reducing agent or on the enzymatic oxidation of glucose to gluconic acid. In the Folin-Wu technique, glucose is determined in a protein-free blood filtrate by reduction of alkaline cupric sulfate and subsequent reaction with phosphomolybdic or arsenomolybdic acid reagent to form a blue complex which can be estimated colorimetrically. The Nelson-Somogyi method uses a protein-free blood filtrate prepared with zinc hydroxide to remove most of the interfering reducing substances.

The presence of a terminal aldehyde in the glucose molecule is the basis of a colorimetric determination with phenolic hydroxyl reagents (phenol in aqueous methyl salicylate or phosphorylated 1,3-dihydroxybenzene) in the presence of strong sulfuric acid and heat.

The *o*-toluidine procedure is a color reaction specific for hexoses—glucose, mannose and galactose. Since aldohexoses other than glucose are normally present in very small concentrations, results obtained by this method approach the true value of glucose. *o*-Toluidine is condensed with glucose in glacial acetic acid to yield a green chromogen by forming an equilibrium mixture of a glycosylamine and Schiff base.

In the preceding techniques, interfering substances such as lactose, galactose and glutathione are measured and the value is reported in the nonspecific term "sugar." Enzymatic determination with glucose oxidase is the only test specific for blood glucose. Blood glucose is converted to gluconic acid and hydrogen peroxide by glucose oxidase; the peroxide is then estimated by iodimetric procedures or by oxidation of a chromogen (*o*-dianisidine or 2,2′-azino[diethylbenzothiazolinesulfonic acid]) in the presence of a peroxidase to form a colored product. Drugs which cause a slight increase in glucose values include ACTH, corticosteroids, D-thyroxine, diazoxide, epinephrine, estrogens, indomethacin, oral contraceptives, lithium carbonate, phenothiazones, phenytoin, thiabendazole and diuretics. Drug interferences with *o*-toluidine methods, which cause a slight increase, include

ascorbic acid, dextran, fructose, galactose, mannose, ribose, xylose and bilirubin.

Another enzymatic procedure uses the hexokinase-catalyzed conversion of glucose to glucose 6-phosphate (G6P), and then to 6-phosphogluconate and NADPH in the presence of NADP and G6P dehydrogenase. The NADPH thus formed is equivalent to the amount of glucose present and is estimated spectrometrically at 340 or 366 nm.

Normal fasting blood-sugar values for adults are 80 to 120 mg/100 mL; true glucose is 65 to 100 mg/100 mL. When the blood-sugar values exceeds 120 (hyperglycemia), diabetes mellitus should be suspected and can be confirmed by evidence of diminished carbohydrate tolerance. The effect of ingested carbohydrate on blood sugar can be determined by the *glucose tolerance test;* 100 g of glucose (1.75 g/kg) in water or a flavored beverage, is administered orally and glucose determinations are performed on blood and urine samples at hourly intervals for 3 hours. Values above 160 at 1 hr and 110 at 2 hours in blood samples are abnormal. The renal threshold for glucose is 180 to 200 mg/100 mL of blood, and, therefore, sugar should not appear in the urine of normal subjects in the tolerance test.

Hyperglycemia and decreased glucose tolerance are seen in diabetes mellitus (to 500 mg/100 mL) and hyperactivity of the adrenal, pituitary and thyroid glands. *Hypoglycemia*, with a blood-sugar value of <60 mg/100 mL and increased glucose tolerance, is encountered in insulin overdose, glucagon deficiencies and hypoactivity of various endocrine glands. Intravenous glucose tolerance studies are used to circumvent defective absorption of glucose in the gastrointestinal tract, eg, in steatorrhea.

Monitoring hemoglobin A_{1c} is another way to follow patients with hyperglycemia. This is more specific for diagnosing diabetes but less sensitive than the glucose tolerance test.[15] Normally, hemoglobin A_{1c} accounts for 3 to 6% of the total hemoglobin while in diabetics it is 6 to 12%. The concentration of Hgb A_{1c} in the blood reflects the patient's carbohydrate status over a period of time, providing a marker for hyperglycemia. *Pancreatic function tests* include studies on IV and oral glucose, glucagon and tolbutamide tolerance. The beta cells of pancreatic islet tissue secrete insulin and the alpha cells secrete glucagon, a substance antagonistic to insulin and having a hyperglycemic effect induced by its glycogenolytic action. In *glucagon tolerance studies* the effect of parenteral administration of glucagon on blood-sugar values is useful in the diagnosis of pancreatic and hepatic function. *Insulin and tolbutamide tolerance studies* are used in the diagnosis of endocrine disorders, differentiation of insulin-resistant diabetics and determination of functional hypoglycemia and islet-cell tumors.

Galactosemia, the presence of galactose (>4.5 mg%) in blood, is usually due to an inborn error of galactose metabolism. Congenital deficiencies in galactokinase or galactose 1-phosphate uridyl transferase result in inadequate galactose metabolism with accumulation of galactose 1-phosphate in the liver. Oral administration of galactose in galactosemia leads to a decrease in blood glucose and an increase in concentrations of galactose in the urine and blood. Galactose is measured by estimation of NADH liberated in the conversion of galactose to galactonolactone in the presence of NAD and galactose dehydrogenase. Deficiencies in intestinal disaccharidases such as lactase will preclude efficient conversion of lactose to galactose and glucose, and oral administration of lactose will cause no increase in blood galactose and usually produce diarrhea. Galactose-loading studies are useful in the diagnosis of toxic or inflammatory conditions of the liver. In hepatic cirrhosis, there is a decrease in the galactose-metabolizing capacity of the liver due to the inhibition of hepatic diphosphogalactose-4-epimerase.

Lactic acid is a product of glucose metabolism; it is converted into pyruvic acid and NADH by lactic dehydrogenase (LDH) in the presence of NAD. Blood lactic acid is estimated by reaction with LDH to form pyruvate and NADH; the NADH level is determined spectrophotometrically at 340 nm and is a function of lactic acid concentration. It is elevated (>20 mg/100 mL) following exercise, anesthesia and certain types of acidosis. The *blood lactate/pyruvate* ratio should be calculated in order to determine the presence of excess lactic acid in the blood in acidosis, thiamine deficiency and decompensated heart disease.

Blood pyruvic acid is determined by the reverse procedure; ie, the conversion of pyruvate to lactate in the presence of LDH and NADH. Normal blood pyruvic acid ranges from 0.6 to 1.3 mg/100 mL by chemical methods and 0.3 to 0.7 mg/100 mL by enzymic procedures.

Nonprotein Nitrogen (NPN) Compounds—These refer to all nitrogen-containing compounds in biological fluids exclusive of protein, including nitrogen from amino acids, low-molecular-weight peptides, urea, nucleotides, uric acid, creatinine, creatine and ammonia. Blood NPN usually is determined by digesting a protein-free blood filtrate with sulfuric acid in the presence of a catalyst (SeO_2) to convert nitrogen to ammonium sulfate (Kjeldahl digestion—see page 444); the excess acid is neutralized and ammonia determined by Nesslerization or reaction with alkaline hypochlorite.

The normal blood NPN is 25 to 45 mg/100 mL (48% urea N, 14% amino acid N, 4% creatine N, 1% creatinine N, 3% uric acid N and 30% residual N). In renal damage, NPN is elevated to values ranging from 60 to 500 mg/100 mL (*azotemia*). As variations in NPN mainly reflect alterations in blood urea nitrogen (BUN), urea determinations are more sensitive and preferred as a guide to kidney function.

The primary pathway of nitrogen metabolism in man is the synthesis of urea from ammonia in the liver and then rapid renal excretion of urea. In renal disease (*nephritis*), the excretion of urea is diminished and blood NPN and BUN are increased. In BUN procedures, *urea* is converted enzymatically to ammonia by urease; the ammonia then is determined by Nesslerization, reaction with phenol-alkaline hypochlorite, aeration into standard acid and subsequent titration or reaction with salicylate-nitroprusside reagent at pH 12 in the presence of alkaline dichloroisocyanurate to form a green chromogen which can be estimated colorimetrically. The ammonia also can be estimated by spectrophotometric determination of NAD produced in the conversion of ammonia and α-ketoglutarate to glutamate by NADH-L-glutamate dehydrogenase. Direct chemical determinations of urea are based on the reaction with 2,3-butanedione in an acid medium (Fearon reaction).

BUN (normal = 5 to 25 mg/100 mL) is increased in chronic and acute nephritis, metallic poisoning and cardiac failure; reduced levels occur in rapid dehydration or following diuresis. In severe liver damage due to diminished urea formation, an increase in blood ammonia and decrease in BUN are observed. Urine urea output (6 to 17 g/day) is an index of *glomerular filtration rate* (*GFR*) and kidney function. Increased dietary protein and gastrointestinal hemorrhage will increase urine urea. Decreases in urea excretion involve either tubular reabsorption or secretion defects.

The *nitrogen balance* represents the balance between nitrogen input or produced (N_{in}) and nitrogen excreted (N_{out}); in normal individuals $N_{in} = N_{out}$. N_{out} is regulated by renal GFR; in renal disease GFR is decreased, $N_{in} > N_{out}$ and BUN is increased. The rate of urinary excretion of parenterally administered dyes (phenolsulfonphthalein), inulin sodium, *p*-aminohippurate and mannitol are sensitive indices of GFR in *renal clearance studies*.

Creatine (methylguanidoacetic acid) and *creatinine* (creatine anhydride) are involved in the physiology of muscle

contraction. Creatine phosphate is an intracellular source of high-energy phosphate bonds via the reaction of ATP and creatine kinase. Creatinine is the waste product of creatine metabolism and is the normally excreted compound.

Serum creatinine is determined by reaction with alkaline picrate to form a red chromogen. These values usually represent 20 to 30% of noncreatinine-interfering substances. Absolute determinations can be made by the absorption of creatinine from protein-free blood filtrates on aluminum silicate prior to the final determination. Drugs causing nephrotoxicity result in a slight increase in creatinine and those which interfere with color formation in the reaction include bromosulfophthalein (BSP), phenolsulfonphthalein (PSP), acetoacetate, ascorbic acid, levodopa, methyldopa, glucose and fructose. Creatine is determined after hydrolytic conversion to creatinine with boiling, aqueous picric or hydrochloric acid.

Renal clearance of endogenous creatinine is related to GFR and is normally 1 to 2 g/day (creatinine coefficient = 20 to 26 mg/kg/24 hr). Normal serum creatinine is 1 to 2 mg/100 mL; creatine 0.2 to 1.0 mg/100 mL. Higher values (5 mg/100 mL) indicate glomerular damage or cardiac insufficiency.

Uric acid is a catabolite of purine metabolism as derived from nucleic acids or nucleotide cofactors. Direct methods for determining uric acid involve the reaction with alkaline phosphotungstic acid to form a "tungsten blue," which is estimated colorimetrically. In another method, alcoholic NaOH is added to a protein-free filtrate to eliminate interfering reducing substances (ascorbic acid, glutathione) prior to the reduction of uric acid with acid copper chelate to form a cupric chromogen complex.

In indirect procedures, uric acid is hydrolyzed by the enzyme uricase; the decrease in absorbance at 290 to 293 nm is a function of the initial concentrations of uric acid. The normal blood value is 1.5 to 6.0 mg/100 mL. It is elevated in renal disease, gout due to increased metabolic pools of uric acid and leukemia as a result of increased turnover of cellular nucleoprotein.

Amino acid determinations in blood are performed by conventional colorimetric ninhydrin techniques or reaction with alkaline β-naphthoquinone-4-sulfonate. Normal plasma values range from 3.9 to 7.8 mg/100 mL. A variety of metabolic disorders may be detected by analyzing for increased levels of specific amino acids in the urine or blood. Total urine amino acids are determined by formol titration; formaldehyde reacts with basic amino groups and thus permits subsequent titration of the acidic groups of the amino acids. Daily excretion of amino acid nitrogen ranges from 100 to 400 mg, constituting 1 to 2% of total urine nitrogen.

The identification and quantitation of specific amino acids in the blood and urine are accomplished by paper, thin-layer (TLC), column and ion-exchange chromatographic and electrophoretic separation of electrolytically desalted blood or urine samples. See Chapter 29.

Abnormal amino acid metabolism (*aminoacidopathies*) usually results in the presence of abnormal quantities of specific amino acids in the urine (aminoaciduria).

The aminoacidurias are divided into two main groups:

1. *Primary overflow aminoaciduria* in which blood amino acids are elevated [phenylketonuria (PKU), maple syrup urine disease (MSUD), tyrosinosis and alkaptonuria].
2. Aminoacidurias characterized by elevated amino acid urine levels with normal blood levels (*transport diseases* with a defect in the kidney tubule—eg, cystinuria—and "no-threshold" aminoaciduria in which the kidney has no mechanism for reabsorbing the amino acid involved—eg, homocystinuria).

PKU, a disease characterized by mental deficiency, is associated with the presence of phenylpyruvic acid in the urine and elevated serum phenylalanine levels due to a hereditary (autosomal recessive) deficiency of hepatic phenylalanine hydroxylase, which converts phenylalanine to tyrosine. The availability of treatment through dietary intake is predicated upon early detection. Many states have passed legislation for mass-screening for PKU in all infants. The Guthrie test is performed by placing filter paper discs impregnated with serum or blood on the surface of an agar culture medium containing β-(2-thienyl)alanine at a concentration sufficient to inhibit the growth of *B subtilis*. *Phenylalanine* will reverse this inhibition and the Bacterial Inhibition Assay (BIA) is a direct measure of this amino acid. Serum phenylalanine determinations also can be performed by estimating the fluorescence of a complex with ninhydrin and copper in the presence of L-leucyl-L-alanine.

MSUD is characterized by the odor of the urine and rapidly is fatal to infants. It is associated with a deficiency in the oxidative decarboxylation of α-keto acids leading to an accumulation of both the keto and amino acids in the blood and urine (valine, leucine, isoleucine). TLC and BIA assays can be used to detect MSUD.

Alkaptonuria is a rare, hereditary disease in which homogentisic acid cannot be metabolized further due to a lack of homogentisic acid oxidase. This causes homogentisic aciduria, ochronosis and arthritis.

In *Hartnup disease*, indole and tryptophane appear in the urine due to defective renal and intestinal absorption of tryptophane. Tryptophane is an intermediary metabolite in the synthesis of serotonin (5-hydroxytryptamine) and 5-hydroxyindole acetic acid (HIAA). Excessive production of *serotonin* and the presence of its *HIAA metabolite* in the urine are associated with metastatic carcinoid tumors. HIAA is measured after removal of interfering keto acids with dinitrophenylhydrazine, extraction and estimation with nitrosonaphthol reagent.

Routine screening tests for congenital metabolic defects and the substance under test in the newborn include PKU (phenylalanine), MSUD (leucine), tyrosinemia (tyrosine), homocystinuria (methionine), histidemia (histidine), valinemia (valine), galactosemia (galactose or galactose uridyltransferase), orotic aciduria (orotidine-1-phosphate decarboxylase), arginosuccinuria (arginosuccinic lyase), hereditary angioneurotic edema (C^1-1-esterase inhibitor) and sickle-cell disease (hemoglobin S).

The analyses for these substances are based on BIA, metabolite bacterial inhibition assay (MIA), enzyme auxotraph bacterial assay (ENZ-Aux), fluorescent spot tests or TLC and electrophoresis.

Proteins—The *plasma proteins* (albumins, globulins and fibrinogen) are involved in nutrition, electrolyte and acid-base balance, transport mechanisms, coagulation, immunity and enzymatic action. *Total plasma proteins* may be determined by Kjeldahl, Nesslerization, specific ion-pair (bromcresol green dye plus albumin) or biuret procedures. The last technique is based on the reaction of —CONH— groups joined by carbon or nitrogen linkages in protein with alkaline copper sulfate to yield the biuret complex which can be estimated colorimetrically. Total protein also can be estimated by specific gravity, refractometric or UV spectrometric methods. These methods are subject to large errors in the presence of a pathology involving increased glucose, lipid, urea or abnormal protein concentrations.

The *albumin-globulin (A/G) ratio* is determined by the biuret method after precipitation of the globulins with a sodium sulfate-sulfite reagent. The normal range is 5.5 to 8.0 g% total protein with an A/G ratio of 1.4 to 2.4. Changes in total protein and A/G ratio occur in kidney and liver disease, hemorrhage, dehydration, rheumatoid arthritis and multiple myeloma. Gastrointestinal albumin loss, as seen in GI bleeding, ulcerative colitis, sprue and enteritis, can be

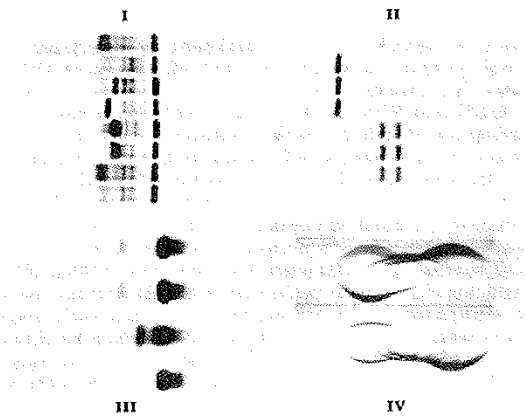


Fig 28-4. Electrophoretic separation of serum proteins (I), isoenzymes (II), hemoglobins (III), and immunoelectrophoresis of plasma protein (IV) (courtesy, Spinco).

detected by monitoring fecal radioactivity after IV injection of ^{51}Cr-human serum albumin.

The physiochemical properties of the plasma proteins—mol wt (68,000 to 300,000) and isoelectric point (pH of minimum solubility and ionic neutrality)—provide the basis for the electrophoretic separation of plasma proteins (Fig 28-4). The plasma sample is spotted on a paper or cellulose acetate strip, or in a polyacrylamide gel (disc or gel electrophoresis) at pH 8.6.

At this pH the proteins are electroanionic and, under the influence of electric current, will migrate to the anode at a rate dependent on their isoelectric point and, in the case of cellulose acetate or gel electrophoresis, their molecular size. The strips are then stained with a protein dye (bromophenol blue, Amidoschwarz or Ponceau S), and the concentrations of the various proteins are estimated by densiometric scanning.

The normal ranges for the major proteins are (in g%): albumin 3.8 to 5.0; total globulin, 2.0 to 3.9; α_1-globulin, 0.1 to 0.5; α_2-globulin, 0.5 to 0.9; β-globulin; 0.5 to 1.2; γ-globulin, 0.7 to 1.6.

Ordinary electrophoresis does not identify the subgroups of *immunoglobulins*, IgA, IgM, IgG and IgE. This is accomplished by immunoelectrophoresis, a process involving electrophoresis and immunodiffusion. The sample is electrophorized in an agar gel (zone electrophoresis) and then antiserum to the specific Ig or to total globulins is placed in a trough aligned parallel to the axis of the original electrophoresis. The serum proteins and antisera diffuse toward each other and form precipitin (antigen-antibody complex) lines. Ordinary cellulose acetate or gel electrophoresis will permit the recognition of diffuse, polyclonal elevation of serum immunoglobulins seen in chronic infections, isolated M-protein peaks of macroglobulinemia and multiple myeloma and absent gamma component in a hypogammaglobulinemia or agammaglobulinemia. Immunoelectrophoresis will indicate specific Ig abnormalities or, by noting the presence of any displacement, bowing or broadening of the precipitin band will aid in the diagnosis of the paraimmunoglobulin monoclonal diseases such as multiple myeloma, macroglobulinemia or chronic lymphatic leukemia.

Radial immunodiffusion is a simple process which also can be used for quantitation of IgA, IgM and IgG.[16] It is performed by incorporating the antibody in an agar gel and then introducing the antigen or test sera into wells punched in the agar. The antigen diffuses radially out of the well into the surrounding gel media, and a visible precipitin line forms where the antigen and antibody have reacted. Quantitation of IgA, IgM and IgG aids in the diagnosis and differentiation of collagen diseases, chronic infections and liver disease. IgE is best quantitated by immunoelectrophoresis or RIA (see section on *Immunology* for the basis and principles of RIA).

Nephelometric techniques detect immunological constituents by measuring the light-scattering properties of various antigen-antibody complexes in a test solution. The Hyland system measures the amount of laser-beam deflection at an angle by employing a photomultiplier tube which is sensitive in the red region of the spectrum. Results are calculated by an electronic-screening system and read in percent relative light-scatter on a digital readout.

Automated electrophoresis equipment offers computer-controlled sample application, staining options, densitometry and pattern interpretation for serum proteins and isoenzymes.

Enzymes—Enzymes are proteins whose biological function is the catalysis of chemical reactions in living systems. Enzymes combine with the substances on which they act (substrates) to form an intermediate enzyme-substrate complex which is then converted to a reaction product and liberated enzyme, which continues its catalytic function. Enzymes are highly specific; a few exhibit absolute specificity and catalyze only one particular reaction, while others are specific for a particular type of chemical bond, functional group or stereoisomeric structure.

Most serum enzymes of clinical significance are intracellular in origin and are elevated in hyperactivity disease, malignancy or injury to cardiac, hepatic, pancreatic, muscle, bone and tissue. As the specific tissue involved will determine the type of enzyme that will be elevated, such determinations are valuable diagnostic tools in the differentiation of various pathological states.

Enzymes are named and classified according to the type of reaction that they catalyze, and to their substrate specificities. Enzyme activity usually is expressed in International Units (IU) where 1 unit (U) is that amount of the enzyme which will catalyze the transformation of 1 μmole of substrate/min at definite temperature, pH and substrate-concentration conditions. Refer to Chapter 52 for a more complete discussion of enzymes.

Transferases are enzymes that catalyze the transfer of amino or phosphate groups from one compound to another. Aspartate aminotransferase (AST) and alanine aminotransferase (ALT) are important in clinical diagnosis. These enzymes catalyze the transfer of the amino group from glutamic acid to keto acids (oxaloacetic or pyruvic) to form aspartic and α-ketoglutaric acids with AST (aspartate aminotransferase) and alanine and α-ketoglutaric acid with ALT (alanine aminotransferase).

Colorimetric methods are based on an estimation of the reaction products (oxaloacetic or pyruvic acid) with dinitrophenylhydrazine, or substrate (α-ketoglutaric acid) by coupling with 6-benzamido-4-methoxy-*m*-toluidinediazonium chloride.

Spectrometric methods are based on the reaction of the product pyruvate with lactic dehydrogenase and NADH, or of oxaloacetate with malic dehydrogenase and NADH. The rate of NADH utilization is measured by the decrease in absorbance at 340 or 360 nm and is directly proportional to transaminase activity.

Normal AST and ALT levels are <40 mU/mL. AST is present in large amounts in liver, cardiac and skeletal muscle, whereas ALT is found primarily in liver tissue. AST is elevated in myocardial infarction and Duchenne muscular dystrophy; AST and ALT are increased in liver disease, acute toxic or viral hepatitis, infectious mononucleosis, obstructive jaundice and hepatic cirrhosis.

Creatine kinase (*CK*) is a transferase found in muscle and brain tissue. It catalyzes the transfer of phosphate groups from creatine phosphate to adenosine diphosphate (ADP) to form adenosine triphosphate (ATP). Activated CK activity is measured by following the increase of ATP in the creatinine phosphate-ADP reaction in the presence of glutathione or cysteine thiol activators. The ATP can be measured by the fluorimetric determination of light emitted by luciferinase conversion of luciferin to adenyl-oxyluciferin in the presence of ATP. Normal serum levels are <50 mU/mL; it is elevated in myocardial infarction and Duchenne muscular dystrophy, but remains at normal levels in liver disease.

Ornithine transcarbamylase (*OTC*) in serum is the only enzyme of the urea cycle which has been used in the clinical investigation of liver disease. It catalyzes the conversion of ornithine to citrulline. The normal serum value is 0 to 0.4 mU/mL.

Oxidoreductases or *dehydrogenases* are enzymes that catalyze hydrogen transfer in cellular oxidation processes. *Lactic* (*LDH*), *α-hydroxybutyric* (*HBDH*), *malic* (*MDH*), *glutamic* (*GLDH*), *isocitric* (*ICDH*) *and sorbitol* (*SDH*) *dehydrogenases* are of diagnostic importance in myocardial and liver disease.

LDH catalyzes the reversible conversion of pyruvic to lactic acid in the presence of NADH. The activity may be estimated colorimetrically by forming the pyruvic acid hydrazone with 2,4-dinitrophenylhydrazine; spectrometric or fluorimetric estimation of NADH in this reaction also is used to estimate enzyme activity. The normal serum LDH value is <200 mU/mL (pyruvate → lactate) and <50 mU/mL (lactate → pyruvate). LDH is increased to a much greater extent and for a more prolonged period than AST or CK in myocardial infarction; it also is increased to varying degrees in certain types of hepatic disease, disseminated malignancies, pernicious anemia and muscular dystrophy.

Recent advances in protein chemistry and technical methodology have led to fractionation of enzymes, previously thought to be homogeneous, into heterogeneous moieties. These multiple-molecular forms of enzymes (*isoenzymes*) have similar substrate specificity but different biophysical properties. LDH, MDH, CK, phosphatases and leucine aminopeptidase exist in isoenzyme forms.

CK isoenzymes are important in the early detection of myocardial damage. Two CK molecular subunits, M and B, produce three isoenzymes: CK-MM found primarily in skeletal muscles, CK-MB in the myocardium and CK-BB primarily from the brain. After acute myocardial infarction (MI), CK-MB appears in the serum in approximately 4 to 6 hours, reaches peak activity at 18 to 24 hours and may disappear within 72 hours. Diagnostic testing of MI includes CK and LDH isoenzymes. Early detection of CK-MB allows the management of myocardial infarcts with agents such as streptokinase or tissue plasminogen activator (TPA). The methods of assessment include electrophoresis, column chromatography and immunoinhibition.

Serum contains five LDH isoenzymes, each a tetramer composed of one or two monomers. LDH 1 and 2 are found in preponderance in heart, kidney and RBC; whereas liver and skeletal muscle largely contain LDH 4 and 5. Intermediate forms prevail in lymphatic tissues and many malignancies. The fractionation of LDH isoenzymes is important in the differential diagnosis of cardiac, muscle and liver disease. It can be accomplished with DEAE-cellulose chromatography, electrophoresis, sulfite or urea inhibition of specific isoenzymes, thermal stability and substrate-concentration requirements.

HBDH reduces α-ketobutyric acid to α-hydroxybutyric acid in the presence of NADH; estimation of the α-keto acid via hydrazone formation or NADH is the basis of activity measurements. The normal serum HBD level is <140 mU/mL; it is elevated in myocardial infarction. LDH 1 is high in HBDH activity. The ratio of total LDH/HBDH often is used in place of LDH isoenzyme determination. Ratios >0.8 are seen in myocardial infarction and <0.6 in acute liver damage.

MDH and *SDH*, in the presence of NAD, catalyze the conversion of malate or sorbitol to oxaloacetate or fructose, respectively. They are of diagnostic value in MI (MDH >48 mU/mL) and acute liver injury (SDH >96 mU/mL).

ICDH oxidizes isocitrate, in the presence of NADP or NAD, to α-ketoglutarate; it is elevated (>5.0 mU/mL) in acute hepatitis.

Hydrolases are enzymes that catalyze the addition of the elements of water across the bond which is cleaved. *Amylases, lipases, phosphatases, 5′-nucleotidase, γ-glutamyltranspeptidase* and *leucine aminopeptidase* are specific examples of clinically important hydrolases.

Salivary and pancreatic *amylases* hydrolyze the substrate starch to maltose and dextrins. Amylase activity can be measured by procedures based on the loss in certain properties of starch as it is hydrolyzed (*amyloclastic*), or by the generation of reducing substances (*saccharogenic*). The amyloclastic methods use the decrease in viscosity and turbidity of hydrolyzed water-soluble starch substrates, or the reaction of starch with iodine as the method of estimation. A newer procedure uses the colorimetric determination of water-soluble dye-dextrin fragments released by amylolytic hydrolysis of a cross-linked, water-insoluble, dye-starch polymer. The saccharogenic methods determine the reaction products (reducing sugars) by a previously described methodology. The normal serum level is 140 mU/mL; elevations are noted in acute pancreatitis, acute abdominal conditions (perforated peptic ulcer, common bile-duct obstruction) and salivary gland disease.

Lipases catalyze the conversion of triglycerides to glycerol and fatty acids. Clinical determinations are based on the titrimetric analysis of fatty acids liberated from an emulsified olive oil substrate, or fluorimetric estimation of fluorescein liberated from a fluorescein fatty acid ester substrate. Serum lipase is increased in pancreatic carcinoma.

Phosphatases catalyze the hydrolysis of orthophosphoric acid esters, and are classified according to the pH of optimal activity into alkaline or acid phosphatases. Activity (alkaline, pH 8 to 10; acid, pH 4 to 6) is measured with phenyl phosphate, glycerophosphate, *p*-nitrophenyl phosphate or thymolphthalein monophosphate substrates. With the latter two chromogenic substrates, the amount of *p*-nitrophenol or thymolphthalein liberated by phosphatase hydrolysis is estimated colorimetrically in an alkaline medium. With a glycerophosphate or phenyl phosphate substrate, the liberated phosphorus is determined by molybdenum blue formation with phosphomolybdic-phosphotungstic acids; phenol also may be estimated with 4-aminoantipyrine or Folin-Ciocalteau reagent.

Acid phosphatase activity may be differentiated by the use of inhibitors in the assay mixture; formaldehyde has no effect on acid phosphatase of prostatic origin, but it inhibits other acid phosphatases, while tartrate is a selective inhibitor of the prostatic enzyme. *Acid phosphatase* is of a primary diagnostic value in metastatic carcinoma of the prostate.

Normal values for *alkaline phosphatase* activity depend on the substrate used; elevations in osteomalacia and in bone tumors depend on the degree of osteolytic or osteoblastic activity. The enzyme (isoenzyme) also is elevated in obstructive jaundice, bone and liver disease.

The enzyme *5′-nucleotidase* is an alkaline phosphomonoesterase that hydrolyzes nucleotides with a phosphate radical attached to the 5′-position of the pentose (eg, adenosine monophosphate). The normal serum value is 17 mU/mL; it is elevated in hepatic disease.

Leucine aminopeptidase (LAP) is an exopeptidase which

hydrolyzes the peptide bond adjacent to a free amino group. It liberates amino acids from the *N*-terminal group of proteins and polypeptides in which the free amino group is a L-leucine residue. Activity is determined by spectrophotometric estimation following hydrolysis of the amide bond of a leucinamide substrate at 238 nm. Clinical estimations usually are performed on synthetic substrates, and since there is no correlation between cleavage of leucinamide and these substrates, the LAP-like activity is designated *leucine arylamidase*. A fluorometric determination of naphthylamine liberated from a leucyl-β-naphthylamide substrate or colorimetric determination of *p*-nitroaniline liberated from leucino-*p*-nitroanilide substrate also has been used. The normal value is 8 to 22 mU/mL; it is elevated in the last trimester of pregnancy, hepato-biliary disease and pancreatic carcinoma.

Serum *γ-glutamyl transpeptidase* (γGT) is increased in diseases of the liver, bile ducts and pancreas. Together with alkaline phosphatase, LAP and 5′-nucleotidase, γGT usually is tested in the group of cholestasis-indicating enzymes. The assay is based on the hydrolysis of γ-glutamyl-*p*-nitroanilide.

Serum lysozyme (muramidase) activity is increased in certain types of leukemia. Serum arginase, an enzyme which hydrolyzes arginine to ornithine and urea, and serum guanase are sensitive indicators of hepatic necrosis.

Lyases are enzymes which split C—C bonds without group transfer. *Aldolase* is a glycolic lyase which catalyzes the reversible splitting of fructose 1,6-diphosphate to form dihydroxyacetone phosphate and glyceraldehyde 3-phosphate. In the estimation of activity, the triose phosphate reaction products are hydrolyzed with alkali and the resultant trioses are reacted with 2,4-dinitrophenylhydrazine to form chromogenic hydrazones for colorimetric analysis. A spectrophotometric estimation is made by coupling the aldolase reaction products with a dehydrogenase acting on one of the triose phosphates and measuring concomitant changes in NADH. The normal value is <8 mU/mL; it is elevated in muscular dystrophy, polymyositis and acute hepatitis.

The significance of serum-enzyme changes in hepatitis is seen in Fig 28-5 and enzyme activity following myocardial infarction in Fig 28-6.

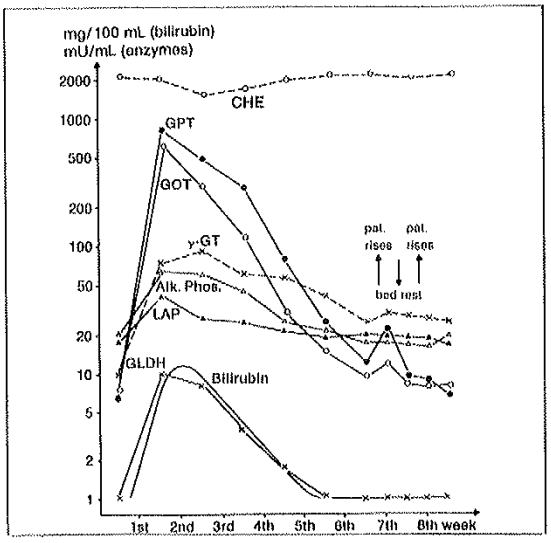


Fig 28-5. Typical course of alterations in serum enzyme activity in acute viral hepatitis (courtesy, Schmidt E, Schmidt FW *Med Welt 21:* 805, 1970).

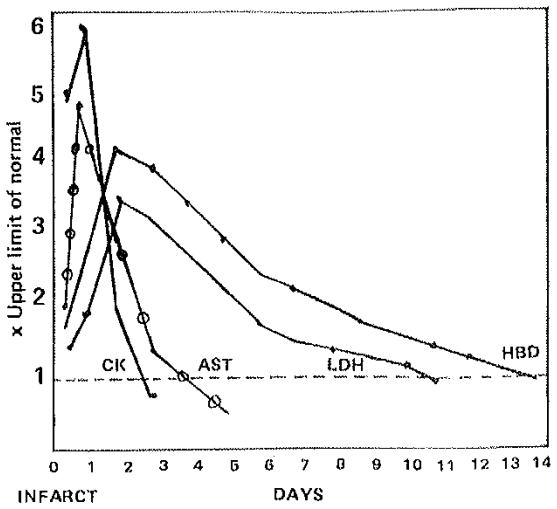


Fig 28-6. Serum enzymes following myocardial infarction, AST, CK, LDH and HBD are compared.

Lipids—The major classes of blood lipids are *fatty acids, cholesterol, triglycerides, phospholipids and lipoproteins.* Hyperlipidemia is not a single aberration and there are a number of different hyperlipidemic states. Lipid-profile tests include measurements of cholesterol, triglyceride, phospholipids and determination of lipoprotein phenotypes.

Cholesterol, a sterol molecule, is an essential substance in steroid-hormone synthesis by the adrenal cortex and bile acid production in the liver. It exists in blood as the free sterol and as cholesterol esters of fatty acids.

In the determination of *total cholesterol*, the serum is extracted with an alcohol-ether mixture and the cholesterol estimated colorimetrically after reaction with acetic anhydride-sulfuric acid reagent (Liebermann-Burchard reaction). The precipitation of free cholesterol with digitonin will differentiate free from esterified cholesterol. Chromatographic separation of cholesterol from its esters on alumina, silicic acid or magnesium silicate columns with organic solvents also has been used.

Gas chromatographic procedures have resulted in the separation and quantitation of cholesterol, its metabolites and precursors; this is a type of partition chromatography in which a volatilized sample is partitioned between a liquid stationary phase and a mobile gas phase. The normal-adult total-serum-cholesterol level is 150 to 270 mg/100 mL; it is increased in hyperlipemia and specifically in hyper-β-lipoproteinemia, nephrosis, diabetes mellitus and myxedema, and decreased in hyperthyroidism and hepatic disease. Free cholesterol comprises 20 to 40% and the ester fraction 60 to 80% of the total serum cholesterol.

Phospholipids are "compound" or "heterolipids" which contain phosphorus, a nitrogen base and a long-chain fatty acid. Lecithin (phosphatidylcholines) and cephalin (phosphatidylethanolamine or serine) are the principal plasma phospholipids, which normally comprise one-third of the total plasma lipids. They usually are bound to lipoproteins. These serum lipids are extracted into an alcohol-ether mixture, digested with sulfuric acid-hydrogen peroxide and the liberated phosphorus determined by colorimetric techniques. The normal lipid phosphorus is 6 to 11 mg/100 mL; about one-half is lecithin. The average ratio of cholesterol to lipid phosphorus when cholesterol is normal is 21. Phospholipid changes usually are associated with cholesterol changes and are of interest in coronary artery and liver diseases and the hyperlipoproteinemias.

Sphingolipids differ from lecithin and cephalin. They are phosphate esters of sphingosine bound to choline or ethanolamine and primarily are found in brain tissue (eg, sphingomyelin, galactolipin). The *ratio of lecithin to sphingomyelin* (L/S) in amniotic fluid or resuscitated amniotic fluid from the oral cavity of the newborn is an accurate assessment of fetal maturity and the respiratory-distress syndrome. Changes in phospholipid biosynthesis during gestation reflect the aging of the fetal lung, as the L/S ratio normally increases.

Tay-Sachs disease is a lipid-storage disease in which the central nervous system degenerates because of the progressive intraneuronal accumulation of excess amounts of the sphingolipid ganglioside GM_2. The accumulation of GM_2 in Tay-Sachs disease has been shown to be caused by a lack of the enzyme hexosaminidase A. Therefore, the measurement of serum, WBC or amniotic fluid *hexosaminidase A* is important in evaluating carriers and in diagnosing Tay-Sachs disease in the fetus.

Both hexosaminidase A (heat-labile) and hexosaminidase B (heat-stable) can catalyze the conversion of 4-methylumbelliferyl-*N*-acetylgalactosamine (a synthetic substrate) to *N*-acetylgalactosamine and 4-methylumbelliferone. The cleavage product, 4-methylumbelliferone, fluoresces under ultraviolet radiation and the intensity of the fluorescence is a measure of the activity of the enzyme. In noncarriers, 50 to 75% of the total hexosaminidase activity is heat-labile (hexosaminidase A), and in carriers 20 to 45% of the total hexosaminidase activity is heat-labile.

The blood fatty acids occur in esterified (EFA) and nonesterified (NEFA) forms. *Triglyceride* determinations are of value in differentiating the hyperlipidemic states, ie, essential (diet-induced) hypertriglyceridemia from familial hypocholesterolemia with or without triglyceridemia. After the preliminary separation from phospholipids, triglycerides most often are determined in terms of their glycerol moiety. The glycerol released by saponification is oxidized to formaldehyde and the latter determined by fluorimetric or colorimetric procedures. Triglycerides also can be determined by coupling the glycerol liberated from lipase/α-chymotrypsin treatment of serum with a glycerol kinase-pyruvate kinase-LDH system and spectrometric estimation of NADH. Normal triglyceride levels are 110 to 140 mg/100 mL. An increase in triglycerides will produce a milky appearance in serum (lipemic). EFA analyses are based also on the reaction of alkaline hydroxylamine with esters of fatty acids to form hydroxamic acids which produce a red color with ferric chloride.

Gas chromatographic procedures have been used to quantitate the various *fatty acids;* ie, palmitic, stearic, oleic, linoleic and linolenic acids. Mono-, di- and triglycerides also can be separated into classes and quantitated by column or thin-layer chromatography, and infrared spectrometry. The total fatty acids of plasma range from 200 to 450 mg/100 mL in the fasting state; they are derived from glycerides, cholesterol esters and phospholipids.

All the lipids in plasma circulate in combination with protein. The free fatty acids are bound to albumin and the lipids aggregate with other proteins to form *lipoproteins*. Electrophoresis and ultracentrifugation are the principal methods used to separate and identify lipoprotein families. *Chylomicrons* (S_f >400), *pre-β-lipoproteins* (S_f 20–400), *β-lipoproteins* (S_f 0–20) and *α-lipoproteins* are the four major classes in order of increasing density and migration on cellulose acetate electrophoresis. Chylomicrons are representative primarily of dietary or exogenous triglycerides, pre-β-lipoproteins of endogenous glycerides, β-lipoproteins of cholesterol and its esters and α-lipoproteins of cholesterol and phospholipids. Abnormal lipoproteins that may appear in plasma include floating β-lipoproteins, lipoprotein X and complexes of normal lipoproteins with IgA and IgG myeloma proteins (autoimmune hyperlipoproteinemia). Age, sex, diet, fasting, posture changes and trauma can alter the lipid profile.

The lipoprotein classes usually are separated by paper, agarose or cellulose acetate electrophoresis. The strips are stained with fat-soluble dyes (Sudan Black or Oil Red O) and quantitated by densiometric scanning. Primary hyperlipoproteinemias are classified into normal and five abnormal types based on cholesterol and triglyceride levels and lipoprotein analysis. Hyperchylomicronemia (Type I), hyper-β-lipoproteinemia (Type II), broad β-band (Type III), hyper-pre-β-lipoproteinemia (Type IV) and hyper-pre-β-lipoproteinemia and chylomicronemia (Type V) are the major classes. Carbohydrate and fat-tolerance studies, post-heparin lipase activity and clinical symptomatology also are integrated into the diagnosis of the various subclasses. The presence or predisposition to coronary artery disease and other disease states is associated with the various types.[17]

Steroids and Other Hormones—The steroids possess a common structure, the perhydrocyclopentanophenanthrene nucleus, and include cholesterol, bile acids, androgens and the adrenocortical, adrenomedullary, estrogenic and progestational hormones.

Androsterone, dehydroepiandrosterone, etiocholan-3α-ol-17-one, 11-ketoandrosterone, 11-ketoetiocholanolone, 11β-hydroxyandrosterone and 11β-hydroxyetiocholanolone are the principal *urinary 17-ketosteroids (17KS)*. These androgenic hormones are derived from the adrenal and, in males, testicular function. The principal urinary steroid metabolites in this group of androgens are found both in the free form, and as conjugates of glucuronides, sulfates or acetates. Their determination in urine involves the acid hydrolysis of the conjugates, extraction with organic solvent, reaction with alkaline *m*-dinitrobenzene (Zimmerman reaction) and colorimetric estimation of the chromogen. The individual 17KS can be separated by TLC prior to analysis to obtain further information on the individual steroids. The normal adult urine values are: male, 9 to 24 mg/day; female, 5 to 17 mg/day. Decreased excretion is seen in hypoactive disease of the pituitary, gonads and adrenals. Increased excretion is seen in hyperplasia, cancer or tumors of the adrenals.

Testosterone is the most potent androgen in blood. The measurement of urinary or serum testosterone is useful in distinguishing normal and hypogonadal males and in treating hirsutism in the female. This hormone is determined by gas chromatography, competitive protein-binding, isotope dilution or RIA procedures. Normal serum testosterone is 0.2 to 1.1 μg/100 mL in the male and <0.1 μg/100 mL in the female.

The natural *estrogenic hormones* are estradiol, estrone and estriol, produced in the gonads, adrenals and placenta. The relative amounts of the three estrogens rise and fall concomitantly during the menstrual cycle. Maternal, urinary total-estrogen excretion, especially estriol, is an indirect index of the integrity and viability of the fetoplacental unit. Analysis involves acid or glucuronidase-arylsulfatase hydrolysis of the conjugates, removal of urinary glucose if present, extraction and colorimetric or fluorimetric analysis. In the determination, after acid hydrolysis and ether extraction of the urine, the estrogens are methylated with dimethyl sulfate and chromatographically separated prior to reaction with phenolsulfuric acid to yield a red chromogen for colorimetric analysis. The normal estrogen output is 4 to 60 μg/24 hr in the female and up to 25 μg in the male. Estrogen deficiency can be related to ovarian failure and pituitary deficiency.

Progesterone is a progestational hormone which is secreted by the corpus luteum of the ovary and also by the adrenal

cortex. Serum progesterone determination is of value in the detection of ovulation and is a measure of the secretory activity of the placenta during pregnancy. Progesterone is determined in serum by RIA, double-isotope derivatization, gas-liquid chromatography or competitive protein-binding techniques. Normal, menstrual-cycle serum progesterone levels vary between 0 and 1.6 μg/100 mL.

Pregnanediol is the principal metabolite of progesterone. The urinary determination of pregnanediol excretion is an indirect index of progesterone levels but is subject to variation due to individual differences in hepatic metabolism of this hormone and is not representative of total endogenous progesterone production.

Adrenal cortex steroids include glucocorticoids, androgens, estrogens, progesterone and mineralocorticoids. Glucocorticoids can be determined as plasma cortisol (plasma 17-OH corticosteroids), urinary-free (unconjugated cortisol) or total-urinary *17-OH corticosteroids*. The latter are determined in urine as *17-ketogenic steroids* (17KGS). The 17KS in urine are reduced with borohydride to alcohols; the 17-OH steroids are oxidized with sodium bismuthate or periodate to 17KS and quantitated by the alkaline dinitrobenzene method. The 17-OH steroids can be quantitated directly by the phenylhydrazine-sulfuric acid reaction after hydrolysis of glucuronide conjugates and chromatographic purification. The 17-OH steroid analysis only determines compounds with the dihydroxyacetone side chain, such as tetrahydrocortisol or tetrahydrocortisone; the 17KGS analysis includes the 17-OH-corticosteroids with the dihydroxyacetone side chain and the pregnanetriol type of compound. Normal 17KGS daily urinary excretion is 5 to 23 mg in the male and 3 to 15 mg in the female. They are reduced significantly in myxedema and adrenal or anterior pituitary insufficiency. Plasma cortisol usually is measured by fluorimetric or gas-chromatographic procedures.

Aldosterone is the most active member of the mineralocorticoid group. The determination of urinary aldosterone is of value in differentiating benign essential hypertension from primary aldosteronism (Conn's syndrome), which is caused by an adrenal adenoma and is accompanied by hypertension. A double-isotope derivatization technique is used. Urinary aldosterone is acetylated with ^{3}H-acetic anhydride; aldosterone-^{14}C-diacetate standard is added early in the procedure. The $^{3}H/^{14}C$ specific activity of the final product is measured after chromatographic purification and is a direct measurement of aldosterone. The normal aldosterone levels of about 10 μg/day are elevated in Conn's disease and usually are associated with low serum potassium, sodium retention and low-concentration alkaline urine.

The anterior pituitary secretes three substances (*gonadotropins*) which regulate gonadal activity: *follicle-stimulating hormone* (*FSH*), *luteinizing hormone* or *interstitial cell hormone* (*ICSH, LH*) and *luteotropin* (*LTH*). The gonadotropins are glycoproteins. Bioassay methods can be used to determine gonadotrophic activity. After fractionation and isolation the urine extract is assayed in test animals as to the follicular growth of the ovaries in hypophysectomized animals or increase in testicular, ovarian or uterine weight in various animal models. RIA techniques have been developed for these gonadotropins and represent the most sensitive and precise measurement method.

Analysis of serum or urinary *placental lactogen* (*HPL*) and *chorionic gonadotropin* (*HCG*), a placental-derived protein hormone, is useful in the diagnosis of threatened abortion, hydatiform mole and choriocarcinoma. HCG, pregnanediol and progesterone as well as total and fractionated estrogens are useful in testing for pregnancy. HCG and HPL readily are measured by RIA and low values are seen in threatened abortion and intrauterine fetal death.

The increase in HCG in the serum or urine of the pregnant female is the basis of a routine *pregnancy test*. Test components consist of an antigen in the form of HCG latex particles and an HCG antiserum. When antiserum is mixed with urine containing a detectable level of HCG, it is neutralized and no agglutination of latex-antigen particles occur (*agglutination inhibition test*). The commercial application of the HCG assay gives laboratories a rapid, accurate pregnancy test by taking advantage of monoclonal antibody specificity and sensitivity. A monoclonal slide procedure on urine, Duoclon (*Organon Diagnostics*), uses two different monoclonal antibodies, one against HCG and one against the HCG_{β} subunit for maximum specificity. Agglutination indicates a positive test with a sensitivity level of 500 miU HCG/mL, detecting pregnancy a few days after conception.

Human growth hormone and *insulin* are proteins which are of diagnostic value in growth-rate studies and diabetes. They are best quantitated by RIA.

Epinephrine and *norepinephrine* are biologically active catecholamines derived from the adrenal medulla and sympathetic nerve endings. Catecholamines are measured in the blood and urine after fractionation on alumina or ion-exchange columns, oxidation at pH 3.5 or 6.0 and subsequent fluorimetric analysis. Urine catecholamines are increased to >350 μg/24 hr in adrenal medullary tissue tumors (pheochromocytoma). The normal plasma level is 2.1 to 6.5 μg/L with about 80% as norepinephrine.

Vanillylmandelic acid (*VMA*) is the urine metabolite of these two catecholamines. Its quantity in urine reflects the endogenous secretion of catecholamines. VMA can be determined colorimetrically, after extraction of the urine with ethyl acetate and diazotization with *p*-nitroaniline and ethanolamine in the presence of carbonate ion. VMA also can be measured spectrometrically following periodate oxidation to vanillin and solvent extraction. The normal output is 0 to 12 mg/24 hr.

Homovanillic acid (*HVA*) is not a metabolite of epinephrine or norepinephrine, but is produced from a common precursor, dopamine. Elevated HVA excretion is diagnostic in cases of neuroblastoma.

The biosynthesis of *serotonin* (5-hydroxytryptamine) and urinary excretion of its metabolite, 5-hydroxyindoleacetic acid (5-HIAA), are increased in argentaffine tumors. These have a very large capacity to metabolize tryptophane stores to serotonin. Urinary 5-HIAA increases from 1 to 7 mg/24 hr to as much as 1 g/24 hr in this type of tumor.

Bilirubin, a tetrapyrrole which is derived from senescent red-cell degradation, normally occurs in low concentration in the blood. In bile, it is present as the water-soluble conjugated acyldiglucuronide. In blood, bilirubin is tightly bound to plasma albumin. The reduction of bilirubin in the intestine yields urobilinogen which is, in turn, oxidized to a brown pigment—urobilin.

Serum bilirubin is determined by coupling with diazotized sulfanilic acid to form azobilirubin for colorimetric analysis. The *direct* or *conjugated bilirubin* test is performed in aqueous media; the *indirect* or *free bilirubin* analysis is performed in methanol or caffeine-sodium benzoate solution. Normal values in serum are: direct, 0 to 0.3 mg/100 mL; total, 0 to 1.5 mg/100 mL.

Clinical jaundice is a yellowing of the tissues associated with hyperbilirubinemia; in hemolytic disease of the newborn due to Rh and ABO incompatibilities, indirect serum bilirubin is elevated, whereas acute hepatitis results in increases in the direct type.

Electrolytes—The normal plasma electrolyte level is 154 mEq/L of cations and 154 mEq/L of anions. The osmotic effects of chloride, bicarbonate, sodium and potassium are important in the maintenance of normal muscle contraction and water distribution between cells, plasma and interstitial fluid.

Flame photometry, atomic-absorption spectrometry, neutron-activation analysis, X-ray fluorescence, ion-specific electrodes and colorimetric techniques are used in the identification and determination of cations or anions in biological fluids. Advances in technology have developed multiphase systems capable of measuring not only sodium and potassium but also chloride, carbon dioxide and calcium simultaneously.

Sodium and *potassium* serum concentrations are readily measured by flame photometry or highly sensitive and specific atomic-absorption spectrometry. The latter technique is similar to emission-flame photometry, except that it measures energy as it is absorbed by atoms rather than as it is emitted by atoms. Both techniques are based on the characteristic absorption or emission wavelengths of the cations. Ion-specific glass electrodes also are used for Na^+ and K^+ determinations, eliminating the use of a flame or combustible gas and can be performed on whole blood, plasma or serum.

Chloride levels in serum or urine are determined by titration with acid mercuric nitrate solution in the presence of s-diphenylcarbazone indicator. They also may be determined potentiometrically with a silver-silver chloride pH electrode assembly. The normal serum values are 135 to 155 mEq Na/L, 3.9 to 5.6 mEq K/L and 95 to 106 mEq Cl/L; urine levels are 150 to 197 mEq Na/day, 20 to 64 mEq K/day and 180 to 270 mEq Cl/day.

Serum sodium, potassium, chloride and *bicarbonate* determinations are useful indicators in adrenal cortical insufficiency, renal and cardiac failure, anuria, dehydration, alimentary tract diseases associated with diarrhea and vomiting and increased renal electrolyte excretion (diuretic therapy).

The determination of excess *chloride* (>50 mEq/L) in the perspiration of patients with pancreatic *cystic fibrosis* is an accurate diagnostic tool. Perspiration is stimulated by placing the patient's hand in a plastic bag for 15 to 20 min or, preferably, by an iontophoresis technique in which pilocarpine nitrate ions are transported through small areas of the skin to produce local perspiration. The chloride content may be quantitated with silver nitrate–potassium chromate-impregnated papers or with ion-selective electrodes.

Bicarbonate, phosphates, sodium, potassium and chloride concentrations are related to maintenance of acid-base balance in the body. The pH of the blood reflects the state of the acid-base balance and is related mathematically to HCO_3^- concentration and partial pressure of CO_2 (pCO_2) in blood by the Henderson-Hasselbach equation.

$$pH = 6.1 + \log \frac{[HCO_3^-]}{[H_2CO_3]} \qquad (2)$$

Blood pH, as measured electrometrically, has a normal range of 7.36 to 7.40 for venous samples and 7.38 to 7.42 for arterial samples. The *pCO_2 level* in blood is determined by measuring the pH of the blood at three different pCO_2 concentrations—one native to the blood and the other two obtained by equilibration with gas mixtures of known pCO_2. Blood bicarbonate levels also may be determined by measuring the amount of acid neutralized by plasma or serum and pCO_2 calculated by Eq 2. The relationship between pCO_2 and carbonic acid concentration is

$$[H_2CO_3] = 0.03 \times pCO_2$$
$$\text{mM per L} \qquad \text{torr} \qquad (3)$$

The role of oxygen and hemoglobin in respiration has been discussed previously. Measurements of blood pH and CO_2 content are used in differentiating respiratory acidosis (low pH, high CO_2) from metabolic acidosis (low pH, low CO_2).

Blood oxygen (pO_2) and *percent oxygen saturation* are measured by a polarographic method; the blood sample is placed in a chamber and separated from a combined platinum and silver-silver chloride electrode by a polypropylene membrane. By diffusion through the membrane, equilibrium is established between the pO_2 of the blood and a film of solution in contact with the electrode. A current, which is proportional to blood pO_2, is generated after the application of a polarizing voltage.

Calcium and phosphorus are important minerals in the processes of bone calcification, nerve irritability, muscle contraction and blood coagulation. Calcium is present in plasma as an ultrafilterable (ionic and nonionic) form and a protein-bound fraction. Blood phosphorus consists of inorganic phosphorus, organic phosphate ester (G6P, ATP) and phospholipids.

Serum and urine calcium levels are determined routinely by titration with EDTA or EGTA using a fluorescent calcein or calcichrome indicator. Other methods are based on the colorimetric analysis of calcium-methylthymol blue complex in the presence of 8-quinolinol to prevent interference by magnesium. Bis-(*o*-hydroxyphenylimino)ethane forms a colored complex with calcium and, in the presence of polyvinylpyrrolidone to inhibit phosphate interference, is a sensitive and specific method for calcium. Calcium is determined best by atomic-absorption spectrometry. As with all cations, calcium can be determined by emission- or absorption-flame photometry or ion-selective electrodes.

Inorganic phosphorus levels are determined by reaction with acid molybdate reagent to form phosphomolybdic acid which, in turn, is reduced with aminonaphtholsulfonic acid or *p*-dimethylaminophenol sulfate to give a blue complex which is estimated colorimetrically. Normal serum levels are 2.5 to 4.5 mg P/100 mL and 9 to 11 mg Ca/100 mL.

Calcium levels are decreased and phosphorus increased in hypoparathyroidism; an opposite effect is seen in hyperactivity of this gland. In rickets and osteomalacia, the concentrations of both elements are decreased. In establishing primary hyperparathyroidism and other causes of hypercalcemia, daily measurements for ionized calcium (Ca^{2+}) are replacing total Ca measurements using ISE technology.

Copper, magnesium, zinc and iron are trace elements in blood. They are quantitated readily by flame photometric, colorimetric or atomic-absorption techniques.

Organ Function Tests—The analyses of various blood or urine constituents, determination of metabolic excretion rates of exogenous compounds or endogenous metabolites and effect of exogenous stimuli on these parameters are used for evaluation of *in situ* activity and function of various organs. Organ function studies are performed in diseases associated with the liver, kidney, parathyroid, thyroid and pituitary gland, gastrointestinal tract, pancreas, adrenals and gonads. The principles and significance of the analysis used in such evaluations have been described also in other sections of this chapter.

Tests for *hepatic function* are based on bilirubin metabolism and excretion, carbohydrate metabolism (galactose tolerance test), plasma-protein changes (cephalin flocculation test and A/G ratio), abnormal fat metabolism, detoxification mechanisms (hippuric acid synthesis), excretion of injected substances [BSP], prothrombin formation and previously discussed enzyme levels.

Diseases of the liver are due to cellular alterations (hepatocellular) or obstructions to the flow of bile (obstructive jaundice). Hepatocellular liver disease can be chronic (postnecrotic cirrhosis, carcinoma) or acute (viral hepatitis, alcoholism, toxin- and chemical-induced).

The *cephalin flocculation test* is based on the flocculation of cephalin-emulsified cholesterol by γ-globulin. In normal serum an albumin-like protein will inhibit this reaction; in hepatic diseases, which produce abnormal γ-globulin or reduced albumin levels, the flocculation will occur.

The *detoxification mechanisms of the liver* can be evalu-

ated by intravenous administration of sodium benzoate and estimation of the benzoic acid metabolite, hippuric acid, in the urine. In hepatoparenchymal disease, a reduced capacity of the liver to form hippuric acid by conjugation of glycine and benzoic acid is observed.

The ability of the liver to excrete an injected dye is determined in the *BSP test;* the serum is analyzed for dye concentration at a suitable time interval after IV administration of 2 to 5 mg BSP/kg. Radioiodinated (^{131}I) Rose Bengal Sodium dye also has been used in dye-excretion studies with isotopic estimation of urine dye levels.

Kidney function tests are based on the determination of blood nonprotein nitrogen (urea, uric acid and creatinine), electrolytes, blood acid-base balance, routine urinalysis and the clearance of administered compounds in the urine. Most *clearance studies* are performed with substances that are not resorbed or secreted by the renal tubules: inulin, mannitol, sodium *p*-aminohippurate or ^{125}I-iothalamate sodium (sodium 5-acetamido-2,4,6-triiodo-*N*-methylisophthalamate). These are administered intravenously and the rate of urine clearance and glomerular filtration is estimated by analysis of the urine. The excretory capacity of the renal tubular epithelium can be determined by measuring the clearance rate of PSP. The dye is injected IV and the rate of its clearance in urine is determined. PSP is bound loosely to serum albumin and is removed rapidly from the blood by the renal tubules.

Sodium iodohippurate-(^{125}I), which is extracted almost completely from the blood on a single passage through the kidney, also has been used in renal function studies; a *renogram* or isotopic scan of both kidneys is performed. The test provides data on renal tubular secretion, renal vascular competence and renal evacuation and is primarily useful as a comparison of individual kidney function. It is important to note that 50% of kidney function can be compromised without any significant change in the routine renal function parameters.

Thyroid function tests usually measure the circulating levels of the thyroid hormones, and not the end-organ effect. The thyroid gland converts inorganic iodide to *thyroxine* (T_4) and *triiodothyronine* (T_3). T_3 and T_4 are stored in the colloid part of the gland as part of the thyroglobulin molecule. Hypothalamic *thyrotropin-releasing hormone* (TRH) mediates the release of the pituitary thyrotropin (*thyroid-stimulating hormone, TSH*). Excess levels of circulating T_4 depress, and low levels of T_4 increase, TSH release. TSH stimulates the proteolytic degradation of thyroglobulin to release T_4 and T_3, and increases organification of iodine. T_4 accounts for 90% of secreted thyroid hormones and exists in blood bound to *thyroxine-binding globulin* (*TBG*) or *thyroxine-binding prealbumin* (*TBPA*) or to albumin. T_3 is not protein-bound and has 5 to 10 times the biological potency of T_4 on a weight basis. Therefore, T_4 represents the major part of protein-bound iodine (PBI). The level of *free thyroxine* (FT_4), the active fraction in blood, is regulated by T_4 and T_3 release and the levels of binding proteins in blood and tissues.

The uptake of orally administered Na ^{131}I preparations by the thyroid gland can be estimated by isotopic scanning of the gland 24 hours after ^{131}I administration and is an index of glandular function (hyperactive, >50% uptake; hypoactive, <15%).

PBI determinations are based on the precipitation of protein-bound thyroxine, removal of inorganic iodine by basic- or anion-exchange chromatography, alkaline incineration to convert thyroxine to inorganic iodide and, finally, quantitation of iodide by reaction with arsenous acid and ceric ammonium sulfate. PBI is a good estimate of total circulating hormonal iodine. The normal range is 4 to 8 μg/100 mL serum.

T_4 can be determined by column chromatography in which it is separated and isolated by ion-exchange chromatography, and then analyzed colorimetrically. Nonisotope thyroid assays have been developed using fluorescence polarization methods for T_4 and free-thyroxin index. In the competitive protein-binding assay for T_4, serum T_4 competes with ^{125}I-T_4 for binding sites on a known amount of TBG. The ratio of bound to free ^{125}I is determined by adsorption of ^{125}I-T_4 not bound to TBG on an anion-exchange resin embedded in a polyurethane sponge or a porous dextran gel, and is a direct index of T_4 levels. The presence of mercurials, inorganic iodide or iodinated radiographic compounds in serum interferes with the T_4 column and PBI procedures. The competitive-binding procedure is affected by the presence of highly protein-bound drugs or changes in TBG levels in serum. The normal range of serum T_4 is 2.9 to 6.4 μg/100 mL by column and 3.0 to 7.0 μg/100 mL by binding assay. T_4 and PBI are increased in hyperthyroidism and the early stages of hepatitis. T_4 and PBI are decreased in hypothyroidism and nephrosis.

FT_4 also is determined in a competitive protein-binding assay in which ^{125}IT_4 and serum are incubated, and then dialysed to determine the percent dialyzable ^{125}I-T_4. FT_4 analysis is used in suspected abnormalities in protein-binding globulins. T_4 binding capacity of serum TBG, albumin and prealbumin can be determined after electrophoretic separation of these proteins.

T_3 analysis is determined by the resin-uptake test. The uptake of ^{125}I-T_3 by a resin is determined in the presence of the test serum. In hyperthyroidism, the primary TBG-binding sites are saturated and ^{125}I-T_3 is taken up by the resin. The resin uptake is decreased in hypothyroidism, and most of ^{125}I-T_3 is bound to TBG in serum. A *free thyroxine index* can be obtained by multiplying T_3 (resin) × T_4 (competitive binding) × 0.01. This product deviates from normal in the same direction as T_3 and T_4 in hyper- and hypothyroidism. This product is stable during euthyroidism in spite of changes in binding proteins; eg, a euthyroid patient on phenytoin therapy will show a decreased TBG and T_4 and increased T_3, but (T_4 × T_3) is normal. The indication of hyper- or hypothyroidism in the presence of abnormal amounts of TBG is observed in the (T_4 × T_3) product.

The determination of *TSH* by RIA appears to be the most useful test in discriminating patients with primary hyperthyroidism from the euthyroidism or hypothyroidism secondary to pituitary disease. Serum TSH is increased in the primary disease state.

The *PBI conversion ratio* is an estimate of the rate of conversion of inorganic iodide to PBI. Radioiodide-(^{131}I) is administered to the subject; after 24 hr, a sample of blood is obtained and the ^{131}I to PB^{131}I is estimated by radiochromatographic procedures with ion-exchange resins (normal conversion, 13 to 42%).

Adrenocortical function is evaluated by estimation of serum or urinary 17-ketosteroids (17-KS) and 17-hydroxycorticosteroids (17-OH-CS) (androgen and corticosteroid metabolism), serum electrolytes (aldosterone metabolism) and blood adrenocorticotrophic hormone (ACTH) levels in the basal state, after stimulation with IM or IV ACTH, or after adrenal inhibition with dexamethasone. In the normal individual, ACTH will increase plasma cortisol and urine 17-OH-CS, and dexamethasone will suppress plasma cortisol. Metapirone, an inhibitor of 11β-hydroxylase, will cause selective secretion of compound S (11-deoxycortisol) by the adrenals in place of cortisol. Compound S will not inhibit the adrenal-pituitary feedback mechanism, the pituitary will secrete more ACTH and the adrenal will secrete more compound S. The determination of urinary 17-OH-CS or tetrahydro-compound S (THS) following metapirone administration is a good index of the functional integrity of the pituitary-adrenal axis; patients with virilizing adrenal hy-

Table IV—Reference Values[a]

Electrolytes		
Calcium	9.0–10.6 mg/dL	2.25–2.65 mmol/L
Chloride		98–109 mmol/L
CO_2 content		23–30 mmol/L
Magnesium	1.2–2.4 mEq/L	0.6–1.2 mmol/L
Phosphorus	2.5–5.0 mg/dL	0.81–1.62 mmol/L
Potassium		3.7–5.3 mmol/L
Sodium		138–146 mmol/L
Metabolites		
Bilirubin	0.1–1.2 mg/dL	1.7–20.5 μmol/L
Cholesterol	150–250 mg/dL	3.9–6.5 mmol/L
Creatinine	0.7–1.5 mg/dL (adults)	62–123 μmol/L
Glucose	60–95 mg/dL	3.33–5.28 mmol/L
Iron	50–165 μg/dL	9.0–29.5 μmol/L
Triglycerides	20–180 mg/dL	0.22–1.98 mmol/L
Urea nitrogen (BUN)	8–26 mg/dL	2.9–9.3 mmol/L
Uric acid	2.5–7.0 mg/dL	0.15–0.41 mmol/L
Proteins and enzymes		
Alanine aminotransferase	(ALT, SGPT)	5–40 U/L at 37°
Albumin	3.5–5.0 g/dL	35–50 g/L
Alkaline phosphatase	35–120 U/L at 37° (adults)	50–400 U/L at 37° (children)
Amylase	60–180 Somogyi Units	110–330 U/L
Aspartate aminotransferase	(AST, SGOT)	8–40 U/L at 37°
Carcinoembryonic antigen (CEA)	<2.5 ng/mL	<2.5 μg/L
Creatine kinase (CK)		10–180 U/L at 37°
Glutamyl transferase (GGT)		5–40 U/L at 37°
Lactate dehydrogenase (LDH)	60–220 U/L at 37°	(lactate → pyruvate)
Total protein	6.0–8.0 g/dL	60–80 g/L
Hormones		
Cortisol in plasma	7–20 μg/dL (at 8:00 AM) (200–550 mmol/L)	3–13 μg/dL (at 4:00 PM) (80–360 mmol/L)
Cortisol (free) in urine	20–90 μg/24 hr	55–248 nmol/24 hr
Follicle-stimulating hormone (FSH)	Adult males	Adult Females
	2–15 mIU/mL	Follicular phase 3–15 mIU/mL
		Ovulatory spike 10–50 mIU/mL
		Luteal Phase 3–15 mIU/mL
		Postmenopause 30–200 mIU/mL
17-Hydroxycorticosteroids in urine	3–10 mg/24 hr	
17-Ketosteroids in urine	5–15 mg/24 hr	(adult females)
	8–20 mg/24 hr	(adult males)
	0.1–3.0 mg/24 hr	(prepubertal children)
Luteinizing hormone (LH)	Adult males	Adult females
	5–25 mIU/mL	Follicular phase 5–30 mIU/mL
		Ovulatory spike 50–150 mIU/mL
		Luteal phase 5–40 mIU/mL
		Postmenopause 30–200 mIU/mL
Metanephrine in urine	<1.3 mg/24 hr	
Prolactin	1–20 ng/mL (males) (1–20 μg/L)	1–25 ng/mL (females) (1–25 μg/L)
Thyroxine (Ty)	5.5–12.5 g/dL (adults) (72–163 nmol/L)	7.8–16.0 μg/dL (newborns) (101–208 nmol/L)
Vanillylmandelic acid (VMA) in urine	<6.8 mg/24 hr	

[a] Serum specimens unless otherwise indicated.[18]

perplasia excrete excessive THS due to a 11β-hydroxylase defect.

Common, chemistry, reference values are listed in Table IV.[18]

Automated Analysis—The automation of analytical techniques used in blood and urine chemistry, hematology, blood typing and immunology has increased the productivity and accuracy of the clinical laboratory.[19] *Computerization* of the automated analytical system also has increased the rapidity of reporting test results, reduced clerical error and provided a unified and updated report of the laboratory tests for each patient.

In the SMA-12 (or SMA-20) Autoanalyzer (*Technicon*), a continuously operating, multiple-channel proportioning pump moves the samples, diluents and reagent streams. Air bubbles segment the flowing streams of samples and reagents, which then may flow through dialyzers to remove interfering substances, move them into chambers preset at desired temperatures and, finally, into detection devices (colorimeters, fluorometers, flame photometers, spectrophotometers). A serum standard is run simultaneously with the samples. The results can be read directly from a recorder or can be coupled into a digital computer output. Sequential, multiple analyses in the SMA-12 are accomplished by distributing the sample to 12 different analytical streams, so that all 12 analyses are in progress at the samo time. The *SMA-12 profile* usually determines calcium, inorganic phosphorus, glucose, BUN, uric acid, cholesterol, total protein, albumin, total bilirubin, alkaline phosphatase, LDH and AST. The Mark X (*Hycel*), Ektachom 400 (*Kodak*), ACA (*Dupont*) and DSA-560 (*Beckman*) also are used in automated clinical-laboratory techniques.

Technicon recently developed the "capsule chemistry" analysis on the Chem 1 analyzer. Microaliquots of the sample (1 μL) and reagents (14 μL) are encapsulated within an inert fluorocarbon liquid. The resulting "test capsule" is introduced into a single, analytical flow path (composed of a solid fluorocarbon liquid, Teflon) where the sample is incubated, mixed, reacted and measured as a moving series of individual tests. The reactions are monitored at in-line detector stations for colorimetric and nephelometric measurements. On each sample 35 chemistries can be run sequentially.

The rapid growth of more-sophisticated chemistry analyzers increases the capacity of any clinical laboratory and is associated with small-specimen requirements incorporating batch analysis, profiles and stat capabilities. In addition to routine chemistry testing, the systems test for enzymes, immunoassay, therapeutic-drug tests, coagulation (fibrinogen, antithrombin III, plasminogen) and electrolytes. Techniques eliminating liquid requirements of other reagent systems are available from *Kodak* and *Ames* using dry reagents,

which are impregnated in pads on a strip or slide and read by a reflectance photometer.

Automated hematology and simultaneous determination of RBC, WBC, hemoglobin and hematocrit, MCV, MCH and MCHC can be performed on the SMA-7A (*Technicon*) Analyzer. The automated Technicon Hemalog system will provide data of SMA-7A and CCV (conductivity cell volume), prothrombin time, partial thromboplastin time and platelet count. *Automated leukocyte differential* was discussed previously under *Hematology*.

Urine

The formation of urine and its excretion are critical physiological activities of the body which provide a mechanism for the maintenance of a constant internal environment for all cells, tissues and organs. This internal ecology of the body is well-recognized and known as homeostasis. Inasmuch as the urine reflects what is occurring within the body, it offers a fluid which is an important source of information that is most useful as an aid in the definition of the states of health and disease. More specifically, the kidney, by means of urine formation

1. Regulates the body water.
2. Excretes metabolic waste products, many of which are of a nitrogenous nature.
3. Excretes toxic substances of both endogenous and exogenous origin.
4. Regulates the electrolyte equilibrium of the body by either excreting or retaining each specific ion.
5. Maintains the delicate balance of pH within the body by excretion of excess acid or excess base.
6. Provides an important route for the elimination of pharmaceutical agents and their breakdown products from the body.

Normal urine contains several thousand compounds most of which occur in minute quantities. Table V identifies some of the constituents of normal urine which are of particular significance.

Urine is studied quite widely as a means of identifying abnormalities associated with disease. The importance of such study is emphasized by the fact that the number of tests carried out on urine far exceeds those made on all other body fluids combined. Urine not only is important in providing information relating to kidney disease, but it may provide information relative to many other body activities. Information from urine studies is of diagnostic value in functional diseases of the kidney, liver, pancreas, blood, bone, muscle and the urinary, gastrointestinal and cardiovascular systems. Urine studies provide vital clinical information on electrolyte and water balance, acid-base equilibrium, intermediary metabolism, inborn errors of metabolism, drug abuse, intoxication, pregnancy and hormone balance. Most of these parameters have been discussed earlier and this section will be devoted to routine urinalysis.

Table V—Normal Constituents of Urine

Constituent	g/day	Constituent	g/day
Water	1400	Amino acids	2.1
Total solids	60	Purine bases	0.01
Urea	30	Phenols	0.03
Uric acid	0.4	Proteins (total)	0.025
Hippuric acid	0.9	Chloride (as NaCl)	12
Creatinine	1.2	Sodium	5
Indican	0.01	Potassium	2
Citric acid	0.8	Calcium	0.2
Lactic acid	0.2	Magnesium	0.15
Oxalic acid	0.03	Sulfur (total)	1.0
Nicotinic acid	0.00025	Phosphate (as P)	1.1
Allantoin	0.04	Ammonia	0.7

It is important to recognize that urine test information, like all other laboratory data, helps provide a picture of the whole body, but any single result requires interpretation to be most meaningful. It also should be recognized that negative results can be essentially as useful as positive results in a great many instances. The ready availability of urine is an advantage that makes it practical as a material for monitoring the course of the treatment of disease as well as for its recognition and definition.

Most urine examinations include observations with regard to the majority of the following—color, odor, turbidity, pH, protein, glucose (or reducing substances), ketone bodies (acetone), occult blood, bilirubin, urobilinogen, bacteria (culture or chemical tests), specific gravity and microscopic examination of sediment, including erythrocytes, leukocytes, casts, epithelial cells, crystals, bacteria, parasites and exfoliative cytology. A *"routine" urinalysis* varies in different institutions but ordinarily involves the inclusion of the majority of the above tests.

Urine for laboratory study should be collected in clean containers—preferably into a disposable unit (polystyrene tube) with a capacity of 15 mL which can be used for collecting, transporting, centrifuging and testing. Refrigeration is desirable for any specimen which is not tested within 1 to 2 hours.

If urine is to be transported through the mails or is to be held for a significant time at room temperature, it is desirable to add a urine preservative (formalin, methenamine, thymol, toluene) which will interfere with microbial growth in the specimen. Several proprietary urine preservative tablets are available. If urine is allowed to stand at room temperature, bacteria will grow in the specimen and cause degradation of many constituents. Frequently, the bacteria decompose urea into ammonium carbonate with a resulting increase in the alkalinity of the specimen. Formed elements, particularly casts and red blood cells, disintegrate in alkaline solution.

The majority of urine tests are done on random specimens but, in certain instances, it is necessary to have a 24-hr specimen for certain specialized analyses. For urine-sugar testing in diabetes detection, it is desirable to use a postprandial urine specimen (ie, after a meal). For protein tests, as well as chemical or culture tests for bacteriuria, the first morning specimen is preferred. Most laboratories use commercially available, standardized, reagent-impregnated strips ("dipstrips") or tablets (*Ames*) for routine urinalysis.

Instrumentation in Urinalysis—Automated urine-testing systems, semiautomated reagent-strip readers and a system which performs the complete urinalysis procedure have been developed. The strip reader is a reflectance photometer which measures urine pH, protein, glucose, ketones, blood, bilirubin, nitrate and urobilinogen. The IRIS AIM (*International Remote Imaging Systems*) measures urine specific gravity by refractometry, urine sediment by staining and classifies analytes, controlled fluid dynamics, video microscopy with an image processor, a chemistry system to read a standard dipstick by reflectance photometry, and color and appearance. These systems achieve standard results for routine urinalysis and increase accuracy and precision.

Volume—The normal volume of urine excreted during a 24-hr period is usually in the range of 1000 to 1500 mL. It is possible for a healthy person to modify the volume either by severe fluid restriction or by ingestion of excessive quantities of fluid. In certain disorders there is a change in urine volume. Urine-volume increases are identified as polyuria and are encountered in diabetes mellitus, diabetes insipidus and in certain stages of chronic renal disease. Urine volume is increased during diuretic therapy and with the ingestion or injection of large volumes of fluid. A decrease in urine

volume usually occurs in dehydration, water restriction and in acute or terminal renal disease. Extensive water loss from severe diarrhea or vomiting causes oliguria or decreased urine volume. Acute renal failure precipitated by shock, poisons or transfusion reaction may result in a complete absence of urine excretion or anuria. In the majority of instances urine study does not require volume measurements, but these are quite critical in severely ill persons where oliguria or anuria is present.

Specific Gravity-Osmolality—The urine density or specific gravity is related to the amount of solids excreted in a given volume of urine. In the majority of instances in healthy persons the specific gravity varies between 1.010 and 1.030 and is related to dietary habits of fluid and food ingestion and, secondarily, to the loss of fluid by other routes such as extensive sweating. The measurement of urine density or specific gravity is a part of "routine urinalysis," and as such provides information with regard to water and solids turnover in the body. The specific-gravity information alone is not nearly so important as it may be in conjunction with other observations. Thus, if dehydration is suspected, a specific gravity in the midrange of 1.015 would cast a doubt about dehydration unless there was a concurrent renal dysfunction.

The kidney possesses a remarkable ability to either form a concentrated urine or a very dilute urine ranging from a specific gravity of 1.001 to 1.032. This concentrating or diluting capacity is diminished in cases where there is a loss of renal function. In fact, one of the sensitive tests for measuring renal function involves the so-called dilution-concentration tests where fluid is administered or withheld, and the specific gravity of the urine is measured. With a serious loss of renal function, the kidney cannot excrete a urine in excess of 1.020 even with marked fluid restriction. In advanced renal disease the specific gravity of the urine may become "fixed" or constant in the range of 1.010 to 1.012 with all urine being of this specific gravity regardless of whether there is overhydration or dehydration.

Specific gravity is measured readily with a special hydrometer, called a urinometer. There is a correlation between the density of urine and its refractive index, and a special refractometer has been designed which gives readings in specific-gravity units on a single drop of urine.

Certain abnormal constituents of urine, such as glucose or protein, when present in high concentrations, will cause significant increases in specific gravity. Certain X-ray contrast media, when excreted in the urine, also will cause marked increases in specific gravity.

Urine specific gravity is only an indirect index of solute concentration, ie, 1 mole of urea will produce a lower specific gravity than 1 mole of glucose. Osmolality is a direct measure of the molal concentration of solutes in solution regardless of their molecular weight, ie, 1 mole of NaCl dissociates into 1 mole of chloride ion and 1 mole of sodium ion. Osmolality is determined in a direct-reading osmometer by comparing the freezing point of urine with that of a standard sodium chloride solution.

The kidneys normally excrete 800 to 1400 mOsm/kg (an osmol is that weight of any substance when dissolved in water depresses the freezing point 1.86°) of solutes per day. Man concentrates urine and eliminates the daily solute load at a maximum volume of 1200 mOsm/kg water. Urine osmolality is an inverse function of urine volume in the normal catabolic state. Urine volume is regulated by the antidiurotic hormone (ADH) and sodium excretion by the hormone aldosterone. Increased osmolality of body fluids stimulates, and increased dilution inhibits, the release of ADH. The major determinant of body-fluid osmolality is sodium. Sodium conservation is mediated through the renin-angiotensin-aldosterone axis. Determinations of plasma and urine sodium, and osmolality and urinary volume, are of diagnostic value in Addison's disease, vasomotor nephropathy (acute tubular necrosis), inapparent volume depletion, incomplete urinary tract obstruction and hepatorenal disease.

pH—Freshly voided urine usually has a slightly acid pH. The normal range is 5 to 8 and, essentially, this is also the abnormal pH range. The kidneys, by reason of excreting a urine of variable pH, provide a regulatory mechanism for the body to get rid of excess acid or alkaline waste products. Since the normal pH range and the abnormal pH range are comparable, the measurement of pH alone provides minimal information, but when used in conjunction with other information, it is a very useful urinary parameter. In conditions of acidosis, the urine is quite acid; in conditions of alkalosis, the urine pH is above 7. When metabolic or respiratory acidosis is suspected, an alkaline-urine pH result almost eliminates the possibility of acidosis. Conversely, if respiratory or metabolic alkalosis is suspected, the excretion of an acid urine indicates that alkalosis is likely not present.

Dip-and-read tests are used widely for pH testing, but pH-meter measurements are used less commonly. In certain situations involving kidney stone susceptibility, it is quite important to maintain a narrow range of urinary pH. For example, in cystinuria an alkaline pH is maintained to keep the cystine solubilized and to avoid as much as possible the crystallization of cystine into renal calculi. The maintenance of urinary pH is also important for optimum results in certain types of drug therapy.

Color—Urine normally has a yellow color, mostly due to urochrome; the color varies from pale straw to dark amber. Darker specimens usually have a high specific gravity. Occasionally, either normal or abnormal urine may show a color different from yellow. Bilirubin may cause fresh urine to be dark in color. In addition, urine which is allowed to stand darkens because of the oxidation of urobilinogen to urobilin. Red, reddish-brown or "smoky" urine usually is due to the presence of hemoglobin (hemoglobinuria), myoglobin (myoglobinuria) or red blood cells (hematuria). Porphyria is an uncommon cause of red coloration. Black urine can be caused by melanin, which may occur in the urine of patients with far-advanced malignant melanoma. An inborn error of metabolism, alkaptonuria, is characterized by the urinary excretion of homogentisic acid, which causes the urine to turn dark brown or black on standing. Many of the unusual colors occasionally found in urine are derived from exogenous sources, including both foods and drugs. Among these are the red color caused by beets, particularly in infants, the golden-yellow or orange-red color of metabolites of pyridium-like drugs or azo drugs and the green or blue color from methylene blue.

Odor—Normal, freshly voided urine has a faint aromatic and characteristic odor, which is more intense in concentrated specimens. If the urine is allowed to stand, the odor becomes strongly ammoniacal and unpleasant because of bacterial destruction of urea. Freshly voided urine having a foul odor indicates severe infection. A sweet, fruity odor may be due to ketones.

Appearance—Freshly voided urine is usually clear. On standing, a precipitate may form which usually consists of amorphous urates if the urine is acid or calcium and magnesium phosphates if the urine is alkaline. The formation of a precipitate is more likely to occur if the urine is refrigerated. Most specimens will become clear again if they are warmed gently to room temperature. Large quantities of mucus, cells, leukocytes or bacteria may cause cloudiness. Protein usually does not cause cloudiness.

Protein—A small amount of protein is present in the urine obtained from healthy subjects although the quantity is not sufficient to give a positive reaction with the tests

commonly used for the recognition of protein in urine. The majority of the 25 to 50 mg of protein that is excreted daily is microprotein (low-molecular-weight polypeptide), with properties quite different than those of albumin and globulin, which are the principal proteins of the blood serum. Albumin and globulins do occur in the normal urine in minute concentrations.

Plasma proteins, hemoglobin, abnormal Bence-Jones protein and proteins (nucleo-, phospho- and glyco-proteins) derived from leukocytes and mucus may be present in urine in nephritis, nephrosis, lesions of the urinary tract, GI dehydration and renal congestion. Abnormal amounts of protein in the urine may be recognized by either precipitation or colorimetric tests. The precipitation depends on the heat coagulation of the protein or on the chemical precipitation of the protein. The most popular of the heat-precipitation tests is the heat-and-acetic acid test in which a tube of urine is heated to boiling after the addition of a drop or two of acetic acid. Sulfosalicylic acid is employed commonly in chemical precipitation tests and, in this test, equal quantities of 3% sulfosalicylic acid and urine are mixed in a test tube and the mixture examined for turbidity indicative of precipitated protein.

Colorimetric tests for proteins involve *dip-and-read* type of systems and are based on the *protein error* of indicators. Certain indicators have a point of color change which is different in the presence of protein compared to the same system in the absence of protein. Thus, by buffering the indicator tetrabromophenol blue on this dip-strip at a specific pH, it is possible to have a yellow color in the absence of protein and a green or blue color in the presence of protein. This test, Albustix (*Ames*), not only indicates the presence or absence of protein in the urine but also can be made to indicate the approximate amount of protein. Strongly alkaline or fermented urines will give false-positive results. The sensitivity of the colorimetric method is such that quantities of 10 to 20 mg of albumin per 100 mL of urine are recognized with confidence.

A positive test for protein in the urine may have any one of several meanings, and it is only when this information is related to other observations that it has optimum value. Proteinuria may be benign and appear following strenuous exercise or simply as a result of standing (orthostatic proteinuria). Protein frequently occurs in the urine during pregnancy and in some instances this is benign, but in other cases it indicates renal complications. Transient proteinuria may occur following severe infections, high fever, exposure to cold and in congestive heart failure. Proteinuria may be an early and sensitive indicator of renal disease and may indicate an abnormality prior to other signs and symptoms of renal impairment in the glomerulus or tubules. In the majority of instances there is not a correlation between the amount of protein in the urine and the severity of the renal disease.

Patients with severe nephrosis may lose up to 25 g of protein per day. Such a marked loss of protein causes a decrease in plasma protein concentration with an accompanying edema. In both chronic and acute glomerulonephritis there is protein in the urine. Tumors of the kidney and renal infection usually will have an accompanying proteinuria. Bence-Jones protein is a unique protein which occurs in the urine of about 50% of patients with multiple myeloma. It has the unusual property of precipitating between 50 and 60° and dissolving at higher temperatures.

Glucose (Reducing Substances)—Glucose normally occurs in urine in such low concentration that it escapes detection by the usual testing methods. The urine of untreated or poorly controlled diabetic patients characteristically contains easily detectable amounts of glucose. A positive test for glucose in urine usually suggests hyperglycemia and the diagnosis of diabetes mellitus; further studies, such as the glucose tolerance test to confirm the diagnosis, are indicated. Glycosuria also may occur when the renal tubules fail to reabsorb glucose normally, and glucose appears in the urine despite normal blood glucose levels, in contrast to true diabetes.

Glucose is the sugar almost always found in urine; however, lactose, galactose, levulose, sucrose and pentoses may be encountered. These other sugars are identified by paper chromatography, selective fermentation, polarimetry, special chemical tests or the formation of their osazones. Other reducing substances occur in urine and may cause falsely positive reducing reactions for glucose. Examples are ascorbic acid, glucuronides, many drugs, homogentisic acid and the preservatives formalin and chloroform.

The traditional test for glucose in urine (Benedict's test) relies on the reduction of cupric ions in alkaline solution to reddish-orange insoluble cuprous oxide. The copper is reduced totally by large amounts of glucose and results in a brick-red sediment with no remaining blue color. Lesser concentrations form green- to rust-colored solutions with some red sediment. A modification of this test, Clinitest (*Ames*), is available in tablet form. The tablet contains copper sulfate, anhydrous sodium hydroxide, citric acid and sodium carbonate. When added to dilute urine, the tablet dissolves and generates enough heat and effervescence to yield results comparable with the Benedict test.

A specific but extremely simple enzyme test for glucose is available—Tes-Tape (*Lilly*), Clinistix (*Ames*) and Multistix (*Ames*). Reagent strips are impregnated with glucose oxidase, peroxidase and orthotolidine. When dipped into a solution of glucose, oxidation occurs and hydrogen peroxide is formed which oxidizes orthotolidine to a blue color. This test is more sensitive than Clinitest, but is not as reliable for estimating the concentration of glucose. The enzymatic test is specific and thus useful in determining whether or not a reducing substance is glucose. Diastix (*Ames*) is a specific urine glucose test using glucose oxidase, which also indicates the quantity of glucose present.

Ketone Bodies—The ketone bodies acetone, acetoacetic acid and beta-hydroxybutyric acid are present in the urine when fats are metabolized incompletely. Ketonuria is seen most commonly in poorly controlled diabetes and indicates ketonemia and diabetic acidosis. Other causes for ketonuria are starvation, fever, protracted vomiting and Von Gierke's disease. Ketonuria also occurs following anesthesia. Acetoacetic acid and acetone produce a distinctive purple color when treated with a mixture of sodium nitroprusside, ammonium sulfate and concentrated ammonium hydroxide. A similar reagent is available in tablet form (Acetest, *Ames*). A drop of urine is placed on the tablet; if ketones are present, a lavender to deep-purple color develops in 30 sec. The color intensity indicates the concentration of ketones. The reagent strip Ketostix (*Ames*), used as a dip-and-read test on urine or serum, contains the same reagents, which are available on Multistix (*Ames*) and other multiple reagents as well. These tests will detect 5 to 10 mg of acetoacetic acid per 100 mL of urine.

Phenylpyruvic Acid—Phenylketonuria (or PKU) is an inborn error of metabolism in which the normal conversion of phenylalanine to tyrosine in the body does not occur and there is a buildup of phenylalanine concentration in the blood. This metabolic disorder causes mental retardation. A portion of the phenylalanine is excreted by the kidneys into the urine and in the process is converted to phenylpyruvic acid (or phenylketone). If this genetic disorder is discovered soon after birth, it is possible to place the infant on a diet very low in phenylalanine-containing proteins and thus minimize the phenylalanine buildup in the body, averting

the serious mental retardation which ordinarily is seen in the untreated PKU patient.

Recognition of PKU can be made by the use of a test for phenylpyruvic acid using a dip-and-read reagent composition containing ferric ions. This test, Phenistix (*Ames*), can be used on urine from all newborn babies. A positive reaction gives a green color, whereas a normal infant's urine gives a pale-ivory or yellow color to the strip. PKU also can be recognized by employing a chemical or microbiological test for elevated phenylalanine in serum, as discussed under *Amino Acids*.

Bilirubin—Bilirubin is found in the urine of patients with hepatitis or obstructive jaundice but not in patients with hemolytic jaundice. Tests for bilirubin and urobilinogen combine to give excellent information in the differential diagnosis of jaundice. Tests for bilirubin are of two kinds; oxidation tests form a green color of biliverdin from bilirubin usually using ferric chloride as the oxidative reagent, and diazotization tests form colored compounds when bilirubin reacts with diazonium salts in a strongly acid medium. Most oxidation tests adsorb the bilirubin onto barium sulfate or similar material before the addition of Fouchet's reagent. The tablet test Ictotest (*Ames*) is the most sensitive diazo test and it uses an absorption mat to concentrate the bilirubin from 5 drops of urine. A reagent tablet is added to the moist spot on the mat and 2 drops of water are added to dissolve the effervescent reagent and wash some of it off the tablet onto the mat where the reaction takes place. A blue or purple color on the mat around the tablet in 30 sec indicates the presence of bilirubin. In addition, a dip-and-read test composition also based on the diazo reaction has been incorporated into the multiple urinalysis reagent strips, Bili-Labstix and Multistix (*Ames*). It is less sensitive than the tablet test, but its convenience allows it to be used in routine urinalysis quite readily. An incidence of approximately 0.1% positives on health-screening population groups, 0.2% on clinic patients and 0.9% on hospitalized patients has been reported.

Urobilinogen—Bilirubin in the bile is reduced to urobilinogen by bacteria in the lower intestine. A portion of the urobilinogen is reabsorbed from the intestine into the blood. A portion of this urobilinogen is excreted into the urine by the kidney and the balance is re-excreted via the bile into the intestine. Although the quantity of urobilinogen in the urine is quite small, it is an important indicator of liver function and red-blood-cell catabolism.

If there is an obstruction to bile flow such as in obstructive jaundice, the amount of urobilinogen formed and reabsorbed into the blood and excreted in the urine is decreased. With impairment of liver function, the excretion of urobilinogen in the bile is decreased, the blood concentration increases and there is a corresponding increase in urinary urobilinogen excretion. Actually, the increase in urinary urobilinogen is one of the most sensitive tests for impaired liver function and this test may indicate an abnormality when all other tests of liver function remain unchanged from normal.

In hemolytic diseases where there is an increased rate of hemoglobin breakdown, the amount of bilirubin formation is increased with a corresponding increase in urobilinogen formation and excretion in the urine. The concentration of urobilinogen in urine can be established by the use of a dip-and-read test which uses the interaction of urobilinogen and *p*-dimethylaminobenzaldehyde (Urobilistix, *Ames*).

Hematuria, Hemoglobinuria and Myoglobinuria—Hematuria refers to a condition in which intact red blood cells appear in the urine. This condition is indicative of a specific defect in the microscopic functional unit (the nephron) of the kidney or it may be indicative of bleeding in the kidney, the ureter, the bladder or the urethra. In the female there may be variable numbers of red blood cells in the urine during menstruation.

Hemoglobinuria is a condition in which free hemoglobin is present in the urine without red blood cells. This may be caused by intravascular hemolysis as a result of a transfusion reaction or by poisoning or toxins. The free hemoglobin in the plasma is excreted by the kidney into the urine. In some situations actual total hemolysis of the red cells occurs after they have entered the urine. This occurs particularly with alkaline urines.

Myoglobin is the red respiratory pigment of muscle. This pigment is quite comparable to hemoglobin in its composition and chemical reactions. Myoglobin may be liberated from muscle cells in certain types of injury and, in such cases, will circulate in the plasma and be excreted in the urine. There are also certain genetic muscle disorders in which myoglobin is lost from the muscles and appears in the plasma and subsequently in the urine.

Chemical tests for red cells, free hemoglobin and myoglobin are based on the peroxidase-like activity of hemoglobin or myoglobin. When a chromogen mixture such as orthotolidine and peroxide is exposed to this peroxidase activity, it will interact rapidly to generate an intense blue color. A dip-and-read solid state system is available which is called Hemastix (*Ames*). This specific composition uses cumene hydroperoxide as the peroxide. The same dip-and-read test for occult blood is incorporated as a component part of multiple, urine dip-and-read tests, eg, Multistix (*Ames*).

Microscopic Examination—Ordinarily, urine contains a number of formed elements or solid structures of microscopic dimensions. These are studied readily by centrifuging 10 to 15 mL of urine, pouring off the supernatant and resuspending the sediment in the drop or so of urine which remains in the tube. This suspension of sediment is placed on a microscope slide and viewed with low-power magnification. Specific structures can be studied with higher magnification. The urinary sediments can be classified into unorganized (chemical substances) and organized (cells and casts) constituents.

In an alkaline urine, amorphous or crystalline ammonium-magnesium phosphates, calcium carbonate or oxalate crystals and ammonium urate may occur normally. Amorphous or crystalline urates, uric acid and calcium oxalates normally are seen in acid urines. The presence of tyrosine, leucine or cystine crystals is associated with various diseases. Chemical crystals are identified by solubility in acid and/or alkali, colorimetric reactions and crystalline structure.

The urine sediment ordinarily contains residues of epithelial cells, crystals and an occasional red or white blood cell. Increased numbers of erythrocytes are seen where there is bleeding into the urinary tract. If the red cells are formed into a red-cell cast, it is suggestive that bleeding has occurred at the glomerular level. An increased number of leukocytes is suggestive of infection and inflammation of the kidney. Casts are microscopic concretions which have the form of a tubule; they have a matrix of precipitated protein and, depending on their appearance, may be identified as hyaline, granular, waxy or red-cell casts. Renal-failure casts are larger and are associated with severe necrosis of the kidney.

Numerous crystals, mucus fibers, bacteria, yeast cells, spermatozoa and parasites (*Trichomonas vaginalis*) may be indentified in the urine sediment. The majority of these crystals do not have any unusual significance but in certain disorders may be indicative of crystal deposits in kidney tissue or predisposition to formation of calculi.

Tissue cells can be recognized in urine sediment. This provides an excellent means of detection and diagnosis of cancer of the lower urinary tract when the sediment is fixed in alcohol and stained by the Papanicolaou procedure. Ex-

foliative cytology of urine may be applied as a routine to all urology patients. In one large clinic the number of positive cases found among urology patients was almost 5%, which is a much higher return of positive results than is obtained with routine staining of cervical smears.

Bacteria—Freshly voided specimens of urine ordinarily contain a few microorganisms, which primarily represent bacteria picked up from the external genitalia. There are fewer contaminating organisms in a *clean-catch* specimen, which involves extensive washing of the external genitalia prior to collection of the specimen. A specimen collected at the midpoint of urination or a "midstream" specimen ordinarily has more organisms than a clean-catch specimen, but fewer than a so-called random specimen. When there is an infection of the kidney or urinary tract, the number of organisms in the urine is increased markedly. Ordinarily, if the urine contains 100,000 or more organisms per mL, the result strongly suggests the presence of an active infection. Infection of the urinary tract with accompanying bacteriuria is relatively common in young girls and women. Quite often the condition is asymptomatic and is recognized only as a result of a study of the urine. If bacteriuria is not treated, it may lead to serious renal injury.

If there is a very large number of bacteria in the urine, the specimen actually may be turbid. This can be recognized by gross visual inspection of the urine. Bacteriuria also can be recognized by microscopic examination of the urine sediment particularly if there is a large number of organisms present. The most widely employed procedure for recognizing bacteria involves plating a specimen of diluted urine on a culture plate, incubating it and counting the number of colonies. A more convenient approach to this same measurement involves the use of a microscope slide which is coated with nutrient agar. Such a slide, when dipped in a urine specimen and then incubated, will indicate the presence or absence of bacteriuria and also the approximate count.

Methods to determine the presence of significant numbers of bacteria in urine samples are available on various automated systems.[20] The Bac-T-Screen (*Marion*) system is a dispensing and filtering system used with a straining process to detect the presence of bacteria on special filter cards by noting the color change on the card. Analysis on the Abbott MS-2 performed by photometric monitoring of bacterial growth changes the light transmitted in a broth culture over a period of time. A decrease in the light transmission due to turbidity or color identifies a positive specimen. The Lumac Biocounter M2010 measures bacterial adenosine triphosphate (ATP) in urine by the bioluminescence produced in a luciferin-luciferase system. Once these rapid techniques are performed to determine which specimens have increased bacteria, further identification and sensitivity testing are performed. Chemical tests for the metabolic activity of bacteria have been used in studying bacteriuria. The most popular chemical test is that for nitrite. Ordinarily, all urine specimens contain nitrate, but do not contain nitrite. If *E coli*, or certain other organisms, are present in sufficient numbers, they will reduce the nitrate to nitrite.

Calculi—Knowledge of the composition of renal and bladder calculi ("stones") is essential in planning the therapeutic regimen for such diseases. Mixed calcium phosphate and oxalate stones usually occur over the entire urine pH range. Uric acid, cystine and calcium hydrogen phosphate calculi generally are associated with acid urines, while magnesium ammonium phosphate calculi usually occur in alkaline urine. Hyperexcretion of one of the calculi components, pH, renal blockage and the presence of foreign objects in the urinary tract are the most probable causal factors in the formation of renal calculi. Calcium oxalate stones are the most common type. The chemical content of the stones is established by routine qualitative analysis for calcium, magnesium, ammonium, phosphate, carbonate, oxalate, uric acid and cystine. Subsequent confirmation by optical crystallography, X-ray diffraction and infrared spectroscopy is also used in the characterization of the physical properties of the calculi.

Feces

Normal feces consists of undigested food remnants, products of digestion, bacteria and secretions of the gastrointestinal tract. *Macroscopic, chemical* and *microscopic* determinations are performed routinely. The normal quantity of feces is about 200 g/day. The brown color is due to the reduction of bilirubin to urobilinogen and then to uribilin (stercobilin); bilirubin is not normally present in feces, but porphyrins and biliverdin (a component of meconium) are excreted during the first days of life. Bilirubin can be detected by tests previously described for bile pigments.

Color changes in the stool can be the result of dietary intake or diagnostic for biliary obstruction and gastrointestinal bleeding.[21] Patients with steatorrhea and malabsorption may show a yellow bulky stool containing fat and gas. The feces is clay colored when bile is prevented from entering the gut. A red or black stool can occur when excessive doses of anticoagulants, phenylbutazone or salicylates are taken, producing bleeding in the gastrointestinal tract. Substances which interfere with the coloration of the stool include antacids (whitish or speckling), bismuth salts (black), iron salts (black), pyridium (orange), senna (yellow to brown) and tetracyclines (red).

Fecal urobilinogen can be determined colorimetrically by reduction of urobilin to urobilinogen with alkaline ferrous sulfate, and then reaction with acidified *p*-dimethylaminobenzaldehyde (Ehrlich's reagent). It is increased from a normal range of 40 to 280 mg a day to 400 to 1400 mg in hemolytic jaundice (dark brown stool), and is decreased in obstructive jaundice (clay-colored stool).

Porphyrins and porphyrinogens do not arise from hemoglobin catabolism, such as bilirubin, but are by-products of the synthesis of heme. Increases in fecal and urinary elimination of coproporphyrin, uroporphyrin and protoporphyrin are valuable diagnostic aids in distinguishing the various hepatic and erythropoietic porphyrias. Fecal coproporphyrins (CP) and coproporphyrinogens (CPP) are determined after extraction, conversion of CPP to CP by iodine and triple-point spectrophotometric estimation at 380, 401 and 430 nm to correct for interfering substances (also see *Urinalysis*).

Fecal occult blood is detected readily by the *o*-tolidine, benzidine, guaiac or diphenylamine tests; this is valid only if the patient has been on a meat-free diet for 3 days. Guaiac and diphenylamine are preferred due to the carcinogenic potential of the other two chemicals.

The Hemoccult test kit (*SmithKline Diagnostics*) uses an impregnated guaiac paper slide for detecting occult blood, which is a useful screening test for colon cancer. Two slides are prepared each day for 3 days from different parts of the same stool while the patient is on a meat-free high-bulk diet. Interfering substances include aspirin, indomethacin and corticosteroids, because they can produce bleeding, and Vitamin C, which interferes with the oxidation reaction of the test. If bleeding occurs high in the GI tract, the blood is digested and converted to acid hematin; 50 mL of blood in the feces will cause melena (black stool). Bleeding from the lower GI tract is apparent from red streaking of stools. The use of ^{51}Cr-tagged erythrocytes has been used to quantitate and locate the source of gastrointestinal bleeding. The subject's red cells are mixed with an isotonic ^{51}Cr solution and then reinjected intravenously. If bleeding occurs, the ^{51}Cr-

isotope content of the feces will be increased. Location of the hemorrhagic area also can be approximated by an isotopic scan of the abdominal area.

The presence of excessive quantities of *mucus* is usually indicative of dysentery, colitis or other inflammatory processes in the intestinal mucosa. Strongly alkaline or acidic reaction in the feces is indicative of excessive quantities of protein or carbohydrate in the diet, respectively.

Quantitative determination of *fecal nitrogen* is useful in analysis of pancreatic function. In pancreatic disease, increases in fecal nitrogen will occur as a result of decreased secretion of pancreatic proteolytic enzymes. The normal individual will excrete 4 to 13% of ingested nitrogen in the feces; in chronic pancreatitis, 9 to 30%. Fecal nitrogen can be determined by the Kjeldahl digestion procedure.

Fecal fat is present in the form of triglycerides of fatty acids (neutral fat), free fatty acids (FFA) and soaps. Fat determinations are based on the solubility of neutral fat and FFA in ether; the soaps are insoluble in ether and have to be acid-hydrolyzed to their respective FFA prior to extraction. Neutral fat will liberate FFA only on alkaline hydrolysis. The FFA, isolated from the above fractionations, are then determined by titrimetric, colorimetric or gas-chromatographic procedures.

Determinations of blood, urine and fecal ^{125}I after oral administration of an iodinated glyceryl trioleate or ^{125}I-oleic acid preparation is an index of *pancreatic, biliary and intestinal absorptive function* and correlates with *fecal fat excretion*. The bile must emulsify the ^{125}I-triglyceride prior to enzymatic hydrolysis by pancreatic lipase to yield FFA-^{125}I, which subsequently is absorbed and metabolized. An increased amount of ^{125}I in the feces is associated with pancreatic diseases (cystic fibrosis with achylia), obstructive jaundice, malabsorption disease (sprue, celiac disease) and steatorrhea. The latter entity can be differentiated as to a pancreatic lipase or intestinal absorptive defect. In the "absorptive" disease, increased excretion of ^{125}I is seen after administration of ^{125}I-triolein or oleic acid. In the pancreatic defect, adequate absorption of ^{125}I oleic acid occurs but fecal ^{125}I is increased after the triolein meal.

A *microscopic examination* of emulsified feces includes analysis for the presence of crystals, food residues, body cells, bacteria and parasites. Crystals of triple phosphate, calcium oxalate, fat and cholesterol, starch granules, vegetable fibers and neutral fat globules are normally present. Octahedral needle-shaped crystals (Charcot-Leyden crystals) are present in parasitic infestation and mucous colitis. Excessive quantities of fat or starch are seen in malabsorption disease.

Adult, larval or ova phases of parasites may be encountered in the feces. The most common parasitic infestations are caused by *cestodes* (tapeworms), *trematodes* (flukes), *nematodes* (roundworms) and *protozoa* (amoeba) (see *Microbiology*).

Toxicology

The determination of drug or chemical concentrations in biological fluids is an important aspect in diagnosing and treating the toxic syndrome induced by various agents in acute or chronic drug-abuse situations or in chemical poisoning.

Barbiturates, glutethimide, methaqualone, chlordiazepoxide, diazepam, diphenhydramine, ethchlorvynol, morphine, phenothiazines and salicylates are encountered in drug-abuse situations. Preliminary screening of serum or urine samples for drug substances is accomplished by TLC procedures. The analysis of serum or urine levels of intact drug or its metabolites usually is performed by extraction of the sample with an organic solvent, separation by gas-liquid (GLC), or high-performance liquid (HPLC) chromatography, and quantitation by spectrophotometric, fluorometric or electrochemical techniques. The interpretation of the serum-concentration data in relation to clinical significance and toxicology must not be limited to numbers.

In acute drug overdosage the time of drug ingestion, time of blood or urine sampling and severity of clinical symptoms or time of death must be interpreted in reference to data on the absorption, tissue distribution, metabolism and elimination of the drug and its metabolites. The specificity of the chemical assay as to interference from other drugs or metabolites of the parent drug must be considered. The combined techniques of GLC or HPLC and mass spectrometry confirms the identity of specific drugs in biological matrices. The extent of absorption of many drug substances is not related directly to the dose when large amounts of a drug are ingested, in comparison to the therapeutic dose.

The tissue-distribution and metabolic rates can be affected by large drug overdoses in which renal or hepatic failure is encountered. The plasma-elimination rate also can be affected, and it is important to recognize the change in elimination kinetics and to be aware of the nature of plasma elimination as defined by a mono-, bi- or polyexponential elimination curve. The drug overdose usually involves several drug substances and the chemical, metabolic and pharmacological aspects of drug interaction must be considered.

The methodology for the analysis of drugs in biological fluids or tissues can be found in the books listed in the *Bibliography*. Analysis for serum *barbiturate* levels will be described in this section as a specific example of the analytical methodology.

Serum is extracted at pH 6.5 with chloroform; the chloroform extract is washed with pH 7.0 phosphate buffer and extracted with 0.45*N* NaOH. The UV spectrum of the alkaline aqueous layer is determined at pH 13 and 10.5. The UV spectra are characteristic and distinguish barbiturates, *N*-methylbarbituric acids and thiobarbiturates. The barbiturates also can be detected by acidifying the alkaline layer, extracting with chloroform and spotting this organic extract on a silica-gel TLC plate. Sequential spraying of the plate with $KMnO_4$, $HgSO_4$ and diphenylcarbazone will show R_f values and color reactions typical of the various barbiturates. Blood barbiturates can be determined more accurately by a GLC procedure in which the retention times are used to identify the specific barbiturates. The degree of severity of clinical symptoms has been correlated with blood barbiturate levels. Comatose, areflexic signs are observed at 5.0 mg% amobarbital, 2.0 mg% pentobarbital, 8.0 mg% phenobarbital and 1.5 mg% secobarbital.

Opiates, amphetamines, barbiturates and methadone can be detected rapidly by *"homogenous" enzyme assay*.[22] In this procedure, the addition of drug antibodies to a conjugate of drug and lysozyme results in the inhibition of lysozyme activity. The addition of free drug to this reaction mixture increases the enzyme activity in proportion to the amount of free drug added. The sensitivity of this type of assay is 0.1 μg/mL of amphetamine and barbiturates, 0.5 μg/mL of methadone, 0.3 μg/mL of opiates and 1.0 μg/mL of benzoylecgonine, a cocaine metabolite. This assay is applicable to large drug-screening programs.

Electron-spin-labeling techniques also can be employed on large-scale drug-screening programs. In this procedure known amounts of drug antibodies are mixed with drug labeled with a stable nitroxide radical (spin-label) and with the specimen to be analyzed. Due to the competition for antibody between spin-labeled drug and drug in the specimen, the spin-labeled drug becomes detached from the antibody and can be detected by electron-spin resonance spectroscopy. This procedure is 1000 times more sensitive than TLC.

Blood-alcohol levels may be determined by aeration, distillation, gas chromatography or specific enzymatic analysis with alcohol dehydrogenase. In the chemical techniques the blood sample is either oxidized or distilled into a dichromate-sulfuric acid mixture; the excess dichromate is then determined by titration with potassium iodide or methyl orange-ferrous sulfate solutions or by colorimetric analysis. The gas-chromatographic and enzyme procedures are specific for ethanol, whereas the chemical techniques are influenced by other volatile or oxidizable substances in the blood. The enzymatic method is based on the reaction of ethanol and NAD in the presence of alcohol dehydrogenase to form acetaldehyde and NADH; the acetaldehyde is removed with semicarbazide and the NADH formed in the reaction is estimated spectrophotometrically at 340 nm. Ethanol levels of >0.10% are indicative of intoxication and apparent psychomotor disturbance. Levels of 0.40 to 0.50% are associated with medullary and diencephalic disturbances such as tremors, coma, respiratory depression, peripheral collapse and death.

Specific analysis of heavy metals is best performed by atomic-absorption spectroscopy. Analyses for arsenic, beryllium, bismuth, copper, iron, lead, lithium, mercury, nickel, thallium and zinc are encountered frequently in the toxicology laboratory. *Blood lead* is determined by forming a lead-dithiocarbamate chelate in the presence of ammonium pyrrolidinedithiocarbamate and extracting the chelate into methyl isobutyl ketone for subsequent atomic-absorption analysis. A lead concentration of >60 μg/mL in children usually reflects significant absorption and accumulation of lead and is interpreted as an indicator of lead toxicity (plumbism).

Increased lead exposure will result in a decrease in delta-*aminolevulinic acid* (*ALA*) conversion to porphobilinogen by ALA-dehydrase in heme synthesis. ALA blood levels will increase to the point that ALA is excreted in the urine. Determination of urinary ALA is performed by removing urine porphobilinogen and urea by ion-exchange chromatography, reacting ALA with *p*-dimethylaminobenzaldehyde and determining the chromogen colorimetrically. Urinary ALA levels >2.5 mg/100 mL are unacceptable in children and industrial lead workers. Urinary ALA levels are not as sensitive an indicator of lead toxicity as blood lead, but they can be used to monitor prophylactic treatment procedures.

Cholinesterase determinations are of value in the diagnosis of suspected cases of organophosphate or carbamate pesticide poisoning. Two types of cholinesterase are found in tissues. True cholinesterase is found in RBC and nerve tissue and exhibits a specificity for acetylcholine substrate. Pseudocholinesterase is found in plasma and has a greater affinity for hydrolyzing butyrylcholine and other esters. The organophosphate and carbamate insecticides inhibit both enzymes. The activity of the plasma enzyme is inhibited more rapidly than the RBC cholinesterase, and recovers more rapidly due to synthesis of new enzyme by the liver. The recovery of the erythrocyte enzyme is slow and is governed by red-cell turnover rate. Cholinesterase activity usually is determined by measuring changes in pH after the incubation of plasma or RBC lysates with acetylcholine. The normal range of this enzyme is 4.5 to 10.9 (plasma), 3.4 to 5.7 (whole blood) and 6 to 10.5 (RBC) units/mL.

Gastric Analysis

The chief constituents of gastric juice are hydrochloric acid, gastric proteases (pepsin and gastricsin), hematopoietic factor (intrinsic factor and vitamin B_{12} binders), gastric hormones and mucosubstances (aminopolysaccharides, mucopolyuronides, mucoids and mucoproteins). Tests for *gastric function*[23] usually are performed on gastric juice samples collected by direct intubation into the stomach. The fasting content (normal, <100 mL) of the stomach is removed and gastric secretion is collected in the basal state, or after stimulation by the oral administration of caffeine-benzoate or alcohol, or parenteral administration of histamine, insulin or the hormone pentagastrin. Samples are collected by continuous aspiration and analyzed for acidity and gastric protease activity at various time intervals. The extent of recovery of total juice can be estimated by oral, nonabsorbable indicators (polyethylene glycol-^{14}C, phenol red and ^{125}I-HSA) instilled into the stomach prior to the aspiration. The recovery and specific concentration of these indicators in gastric juice is an index of gastric secretory volume, completeness of collection and gastric emptying rate.

Gastric juice is a heterogeneous mixture of clear juice and flocculent, clear mucus. The *color* of the juice should be noted as to the appearance of blood, bile and excessive quantities of mucus. The *acidity* can be determined by a simple pH measurement and conversion to mEq H^+ or by titration of centrifuged gastric juice to pH 3.5, 4.5 and 7.4, the respective end-points for free acid (HCl), protease activity and physiological neutrality. The *basal acid output* is about 1 mEq/hr in normal subjects and 2 to 4 mEq/hr in duodenal ulcer patients. The *peak acid output* (*PAO*) after histamine stimulation is 10 to 20 mEq/hr in normals and 40 to 50 mEq/hr in duodenal ulcer; PAO following pentagastric stimulation is similar to histamine. Gastric acid secretion is decreased in atrophic gastritis, gastric carcinoma and certain types of gastric ulcer. Hypersecretion is seen in duodenal ulcer, Zollinger-Ellison (ZE) syndrome and hyperparathyroidism.

In situ measurements of pH may be made with a *Heidelberg capsule apparatus*. In this technique the subject swallows a small pH-sensitive capsule (transmitter); radiowaves are transmitted from the capsule to a sensing device (receiver), and the signals are recorded as a function of pH. The normal pH of the stomach is 1.2 to 1.8.

Tubeless gastric acidity analysis is performed by oral administration of Diagnex Blue (*Squibb*), a carbacrylic ion-exchange resin reacted with azure blue dye. The hydrogen ions in the gastric juice exchange with the dye on the resin; the dye is absorbed and then excreted in the urine. The dye concentration in the urine is a function of gastric acidity. The normal value is >0.6 mg of dye in the urine 2 hours after administration.

The principal gastric proteases are *pepsin and gastricsin;* pepsinogen is a precursor which is converted to active pepsin by free HCl and by an autocatalytic process. *Total gastric protease activity* is determined on hemoglobin or radioiodinated human serum albumin (RISA) substrates at pH 1.8 to 3.1 (RISA-^{125}I); protease activity on hemoglobin will liberate tyrosine which can be estimated spectrophotometrically at 280 nm; with RISA, liberated tyrosine-^{125}I, as estimated by isotopic procedures, is an index of proteolytic activity.

Pepsin activity can be distinguished from the total protease activity by estimation of the 3,5-diiodotyrosine liberated from *N*-acetyl-L-phenylalanyl-3,5-diiodotyrosine substrate at pH 2.1. Pepsin will react on this substrate; gastricsin will not. Normal gastric juice protease activity ranges from 200 to 1200 μg total protease activity/mL and 50 to 300 μg pepsin/mL. The presence of bile, blood, saliva or excess mucus in the sample will decrease both acidity and gastric protease activity.

Gastrin, cholecystokinin, secretin and pancreozymin are gastrointestinal hormones.[24] The role of gastrin and its interaction with other gastrointestinal hormones in the etiology and proliferation of ulcer disease is of recent interest. Accurate RIA techniques have been developed for gastrin and secretin-6-tyrosine due to the availability of a pure synthetic polypeptide. Biological assays based on the effect of

these substances on gastric, pancreatic and biliary secretion also have been used.

Gastrin is found in various species in two forms, G-I and G-II. The only difference is in sulfation of the 12-tyrosyl residue in G-II of the heptadecapeptide amides. Gastrin is found primarily in the gastrin-producing cells (G-cells) of the antral mucosa. The C-terminal tetrapeptide represents the biologically active part of the molecule. Gastrin infusion will stimulate secretion of gastric acid, pepsin and intrinsic factor. It has a slight secretin-like effect and a powerful pancreozymin-like effect on pancreatic secretion. Gastrin also stimulates bile flow. The instillation of HCl into the stomach will inhibit gastrin release; protein and meal stimulation will increase serum gastrin.

The RIA of serum gastrin is of diagnostic value in the ZE syndrome, pernicious anemia and duodenal ulcer. Basal serum gastrin levels in the normal individual are 20 to 30 μg/mL and increase about 2-fold after a protein meal stimulus. Basal serum gastrin levels in duodenal ulcer are normal or slightly elevated, but increase 4- to 5-fold after a protein-meal stimulus. Basal serum gastrin levels are elevated in ZE to 500 to 4000 pg/mL due to the presence of a gastrin-producing tumor. The ZE patient is uniquely sensitive to IV calcium stimulation which will increase both gastric acid secretion and serum gastrin in this syndrome. Basal serum gastrin levels also are elevated in gastric hyposecretion as seen in pernicious anemia and Type A gastritis, and in chronic renal failure due to the decreased metabolic turnover of gastrin in the kidney.

The RIA of serum gastrin is based on the competition of gastrin in test sample with ^{125}I-gastrin for gastrin antibody binding sites. The antibodies used in this procedure are usually cospecific for G-I and G-II. However, they detect all forms of circulating gastrin, ie, Big-Big Gastrin (G-39), Big Gastrin (mol wt 7000; G-33), gastrin heptadecapeptide (G-17, mol wt 2200), G-13 and G-8 (mini-gastrin). The Big components can be converted to gastrin by trypsin hydrolysis. The significance of changes in the ratio of the circulating gastrins is not known, but it has been suggested that G-39 and G-33 predominate in the basal state and cleave to G-17, which is the major serum form after a protein meal.

Other Body Fluids

Physical, chemical and microscopic examination of cerebrospinal fluid, synovial fluid, human milk, transudates and exudates also are performed by the clinical laboratory. The principles of the various determinations are similar to those described for blood and urine.

Microbiology

Clinical medical microbiology is a science which is concerned with the isolation and identification of disease-producing microorganisms, ie, bacteria, fungi (including yeast), viruses, rickettsia and parasites. The techniques employed in the isolation and identification of the suspect organism(s) involve the propagation on suitable primary culture media, selective isolation on special culture media, use of suitable living host material (mouse, embryonated egg, tissue culture, etc), determination of morphological and, where applicable, staining characteristics of the organism and confirmation by biochemical and/or immunochemical analysis. Suitable animal inoculation, where applicable, may be employed to determine pathogenicity. Site, timing, technique (aseptic), instrumentation, and transportation of clinical specimens (blood, urine, feces, cerebrospinal fluid, etc) are prime variables involved in the final differentiation and confirmation process.

Rapid manual enzymatic and immunological test kits have been introduced to identify pathogens for cerebrospinal fluid analysis. The latex-agglutination test coats a specific antibody onto latex particles and when an antigen is present, the latex particles are visible.[25] In the coagglutination test, the specific antibody is bound to protein A on the surface of a staphylococcal cell and the presence of antigen produces agglutination.[25]

Staphylococcus aureus (*Micrococcus pyogenes* var *aureus*) is a Gram-positive coccus frequently found on normal human skin and mucous membranes and frequently associated with abscesses, septicemia, endocarditis and osteomyelitis. Some strains elaborate an exotoxin capable of causing food poisoning. The primary isolation is on blood agar and in thioglycollate broth. With feces and other heavily contaminated specimens, phenylethyl alcohol agar and/or mannitol-salt agar should be inoculated to suppress growth of other bacteria. The identification of pathogenic staphylococci is based on colonial (pigmentation) and microscopic morphology (grape-like clusters), positive catalase production, positive coagulase production (staphylocoagulase-plasma clotting factor) and positive mannitol fermentation.

Streptococcus pyogenes is another Gram-positive coccus frequently associated with tonsillitis or pharyngitis, erysipelas, pyoderma and endocarditis. Neopeptone agar containing 5% defibrinated sheep blood is preferred for primary isolation and to demonstrate characteristic hemolysin production by observing a zone of clear (beta) hemolysis around the colonies on blood agar. Streptococcal groups are identified by precipitin tests with group-specific antisera for A, B, C, D, F and G. Streptex (*Wellcome Diagnostics*) uses a latex agglutination system for identifying the Lancefield group of streptococci. Other groups usually are not associated with human clinical materials.

Legionella pneumophilia identification includes specimen cultures on lung tissue or sterile body fluids (eg, pleural fluid or pericardial fluid). Direct fluorescent antibody method is a test for *L pneumophilia*. Organisms are best seen in the acute stage of the disease. Since the antiserum is species-specific, polyvalent antisera are necessary for identification.

Neisseria gonorrheae is a Gram-negative diplococcus associated with the venereal disease gonorrhea. The identification is based on the primary isolation of the gonococcus from urethral exudates on chocolate agar or Thayer-Martin (TM) medium. The microscopic observation of Gram-negative intracellular diplococci resembling the gonococcus constitutes a presumptively positive diagnosis of gonorrhea. Confirmation of the oxidase enzyme activity of the gonococci is performed by a reaction with *p*-dimethylaminoaniline which turns oxidase-positive colonies black. A positive oxidase test by Gram-negative diplococci isolated on TM medium constitutes a presumptively positive test for *N gonorrheae*. Final identification rests on typical sugar fermentation or specific (fluorescent antibody) staining.

Neisseria meningitidis is the primary cause of bacterial meningitis and septicemia. The primary isolation is based on culturing of a specimen (blood, spinal fluid or nasopharyngeal secretions) on a Mueller-Hinton medium or chocolate agar containing a vancomycin-colistimethate-nystatin antibiotic mixture. The confirmation of the isolate by biochemical reactions (positive oxidase, positive catalase, etc) and serological agglutination with group-specific (A, B and C) antiserum is used in the differentiation. Young cultures of groups A and C may show capsular swelling (Quellung reaction) in the presence of a specific antiserum.

The enteric bacilli (*Enterobacteriaceae*) are Gram-negative, nonsporulating rods associated with dysentery (*Shigella sp*), typhoid fever (*Salmonella typhi*), urinary tract and tissue infections (*Escherichia coli*, *Proteus sp* and *Pseudomonas sp*), and pulmonary infections (*Klebsiella sp*). The primary isolation of enteric bacilli is on selective and differential infusion agar such as MacConkey and eosin-methyl-

ene blue (EMB), and enrichment media such as selenite broth and tetrathionate broth. The primary isolation of *Salmonella sp.* is on Leifson's deoxycholate citrate agar (LDC) or *Salmonella-shigella* agar (SS); if *Salmonella typhi* is suspected, brilliant green agar (BG) and bismuth sulfite agar (BS) may be used and would constitute a presumptively positive diagnosis of *S typhi*.

The confirmation and identification of enteric bacilli may be performed by serological tests and biochemical reactions: H_2S production (triple-sugar iron agar), indole production, acetylmethylcarbinol production, citrate utilization, urease, lysine and arginine decarboxylase and phenylalanine deaminase activity. Enterotube (*Roche Diagnostics*) employs conventional media to perform 11 standard biochemical tests which can be inoculated simultaneously in one compartmented tube, with a single bacterial colony. The serological identification of *Salmonella* and *Shigella sp* is based on the agglutination of antigens that fall into three categories: "K" capsular (*Klebsiella sp* and *Shigella sp*), "O" (*Salmonella sp, Arizona sp, E coli, Shigella sp*, etc) and "H" flagellar (*Salmonella sp*).

Other Gram-negative rods of medical importance are the hemophilic bacilli (*Bordetella pertussis*, whooping cough and *Hemophilus influenzae*, bacterial meningitis), the hemorrhagic bacilli (*Pasteurella pestis*, bubonic plague, and *P tularensis*, tularemia) and pyrogenic bacillus (*Brucella melitensis*, undulant fever).

Spore-forming Gram-positive rods of medical importance belong to the genus *Clostridium*, which are associated with tetanus (*Cl tetani*), gas gangrene (*Cl perfringens* or *welchii*) and botulism (*Cl botulinum*). The isolation of these organisms requires anaerobic conditions. Once the strain to be identified is obtained in pure culture by single-colony selection, its morphological characteristics are noted; the strain then is grown in a variety of definitive media to determine catalase activity, hydrogen peroxide decomposition and fermentation or hydrolysis of carbohydrates and organic acids. The analysis of fermentation products (gas chromatography) also is used for the identification of pathogenic anaerobic *Clostridia*. The major clostridial exotoxin type can be determined by typing with specific antitoxin sera. A Gram-positive, aerobic, spore-former of medical importance is *Bacillus anthracis*, responsible for anthrax, a disease of animals transmissible to man.

The mycobacteria are acid-fast bacilli associated with tuberculosis in man (*Mycobacterium tuberculosis*), in cattle (*Mycobacterium bovis*) and leprosy (*Mycobacterium leprae*). Tubercle bacilli in man are isolated from sputum cultured on a tubed or bottled egg medium (Lowenstein-Jensen) following enzymatic digestion and concentration of the specimens. A provisional diagnosis of tuberculosis usually is made by demonstrating acid-fast bacilli microscopically, X-ray diagnosis and a positive tuberculin skin test.

Other weakly and partially acid-fast bacilli of medical importance are members of the *Actinomycetales*, *Nocardia asteroides* and *Nocardia brasiliensis*, which are responsible for severe pulmonary infections and cutaneous and subcutaneous abscesses.

Bacteriophages (phages) are a special group of viruses that are hosted by bacteria. Any given phage is highly host-specific and when in contact, lysis of the host occurs (phage-typing). They are used primarily as epidemiological tools in subtyping strains of *E coli*, staphylococci or *Salmonella sp* that are presumed to be related epidemiologically. Phages also furnish ideal material for studying host-parasite relationships and virus multiplication.

The medically important fungal diseases include the superficial mycoses, ie, fungal invasion is restricted to the outermost layers of the skin or to the hair shafts (*Microsporum audouini*, ringworm of the scalp, *Trichophyton sp*, athlete's foot and *Epidermophyton floccosum*, *Tinea pedis*) and the systemic pathogenic fungi (*Blastomyces dermatitidis*, *Coccidioides immitis*, *Histoplasma capsulatum*, *Candida albicans*). The diagnosis of the causative agent is based on the isolation of organisms on Sabouraud's dextrose agar or trypticase soy agar with or without cycloheximide and chloramphenicol to suppress the growth of saprophytic fungi and bacteria, macroscopic examination of morphological characteristics and microscopic examination using KOH or lactophenol cotton-blue stain. Biochemical reactions usually are limited to *Candida sp*. Immunological reactions include skin tests, where applicable, agglutination tests, such as latex particle agglutination for histoplasmosis and tube precipitin and complement-fixation tests.

An *antimicrobial susceptibility test* is a determination of the least amount of an antimicrobial chemotherapeutic agent that will inhibit the growth of a microorganism *in vitro*, using a tube-dilution method, agar cup or disk-diffusion method. The test may function as an aid in the selection of a chemotherapeutic agent by the physician. Also, the concentration of antimicrobial agents in body fluids may be determined by biological assay with an organism of known susceptibility for the specific agent.

The laboratory diagnosis of *viral infections* is based upon (1) examination of the infected tissues for pathognomonic changes or for the presence of viral material; (2) isolation and identification of the viral agent; (3) demonstration of a significant increase in antibody titer to a given virus during the course of the illness; (4) detection of viral antigens in lesions—using fluorescein-labeled antibodies and (5) electron microscopic examination of vesicular fluids or tissue extracts. Blood is used for serological tests but seldom for virus isolation. Acute and convalescent-phase blood specimens must be examined in parallel to determine whether or not antibodies have appeared or increased in titer during the course of the disease. Some examples of human viral infections are respiratory infections (Adenovirus group); diseases of the nervous system, ie, polio and coxsackie viruses of the picornavirus group; smallpox (poxvirus group); measles (paramyxovirus group); chicken pox (herpesvirus group) and influenza (myxovirus group).

Members of *Mycoplasmatacea* pleuropneumonia-like organisms (PPLO) are of a range of size similar to the larger viruses. They are highly pleomorphic because they lack a rigid cell wall, they can reproduce in cell-free media and they do not revert to or from bacterial parental forms as the L-forms. Specimens (sputum, bronchial secretions, urinary sediment, etc) for the primary isolation of mycoplasmas (*M pneumoniae*, *M hominis*, etc) should be cultured on agar media containing peptone, serum, ascitic fluid, whole blood or egg yolk. The species identification may be by growth inhibition on agar medium containing type-specific rabbit antisera. Antigenic variants or subspecies may be detected by immunodiffusion. Various PPLO are pathogenic, parasitic or saprophytic. Mycoplasmas have a predilection for mucous membranes and are associated with primary atypical pneumonia and bronchitis.

Clinical parasitology is a science which is concerned with the parasitic protozoa (amoeba), the helminths (cestodes, tapeworms; trematodes, flukes; nematodes, roundworms) and the arthropods. The identification of protozoan ova is based on detailed microscopic morphology (nuclei, etc) using wet mounts (saline or iodine) or stained preparations (iron hematoxylin, etc) obtained from fecal specimens (fresh or preserved with polyvinyl alcohol), which are concentrated by sedimentation, centrifugation or flotation techniques. Trophozoite and/or cystic stages may be detected in fecal specimens associated with intestinal protozoa as in amoebic dysentery caused by *Entamoeba histolytica*.

The commonly encountered helminths are *Necator americanus* (hookworm), *Trichuris trichiura* (whipworm) and *Enterobius vermicularis* (pinworm); they are identified by

characteristic ova. Characterization of tapeworm segments (proglottids) or head (scolex) in a fecal specimen will differentiate *Taenia saginata* (beef tapeworm) from *Taenia solium* (pork tapeworm). Eggs of *T solium* and *T saginata* cannot be differentiated on a morphological basis.

Adult flukes oviposit a characteristic egg which may reach the urine, sputum or feces. *Schistosoma japonicum* eggs have a small, indistinct spine; *S mansoni*, a distinct, large, lateral spine; and *S haematobium*, a distinct terminal spine.

Arthropoda constitute the largest of the animal phyla which are characterized by a segmented body with the segments usually grouped in two or three distinct body regions, by a chitinous exoskeleton, several pairs of jointed appendages and characteristic internal organs. Most arthropods can be preserved in 70% alcohol. They are of medical importance since they can infest man and cause mechanical trauma or produce hypersensitivity from repeated exposure (*Cimex lectularius*, the bedbug) or by toxin injection (*Latrodectus mactans*, the black widow spider), by skin invasion (*Sarcoptes scabiei*, the itch mite) and by transmitting disease (*Anopheles* mosquitoes, malaria), and *Yersinia pestis* in fleas (plague).

The serodiagnosis of parasitic diseases includes the following immunodiagnostic tests: complement-fixation (trichinosis), precipitin test (schistosomiasis), bentonite flocculation (ascariasis), hemagglutination (echinococcosis), latex agglutination (trichinosis), cholesterol flocculation (schistosomiasis), fluorescent antibody (malaria) and methylene blue dye test (toxoplasmosis).

Immunochemistry

Clinical immunopathology[26] includes *general immunology* (immunofluorescence, immunodiffusion, immunoelectrophoresis and agglutination tests), *radioimmunoassay* (RIA-hormones, vitamins, drugs, immunoglobulins), *tissue typing* (histocompatibility tests in organ transplants), *cellular immunology*, *cancer immunology* and *immunohematology*. Examples of each of these disciplines are discussed in this section and other parts of this chapter.

The ELISA, *enzyme-linked immunosorbent assay*, detects antibodies by an indirect technique using enzyme-linked antibodies to label antigenic substances in tissue or body fluid. The antigen is attached to a solid matrix and reacts with a specimen that may contain a complimentary antibody. The antihuman globulin, which is conjugated with the enzyme, is added and the antigen reacts with the bound antibody of the patient. By adding the substrate molecule the enzyme is detected. This system has been used to identify antibodies to viruses, parasites, bacterial products and quantitation of some drugs.

Antibody response is a complex process involving the lymphoid cell system response to foreign stimulus or antigen. Hematopoietic cells in the fetal yolk sac, liver or marrow develop into lymphoid stem cells which, in turn, differentiate into T-lymphocytes of thymic origin and B-lymphocytes of bone-marrow origin. The T-cells further differentiate into lymphoblasts which are responsible for *cell-mediated cellular immunity* (graft vs host reaction, tissue transplant rejection, tuberculin skin testing, *delayed-type hypersensitivity*). B-cells differentiate into plasma cells which are responsible for humoral immunity which is mediated by circulating serum immunoglobulins (*immediate-type hypersensitivity*).

Macrophages can cooperate in presentation of antigen to the T- or B-lymphoblasts. Cooperation between T- and B-cells, immunological memory, development of immune tolerance to antigens and genetic control of the immune response are integral properties of the immune system and are related to development of immune deficiency and autoimmune disease.

The identification and determination of *immunoglobulins* (IgG, IgM, IgA) by radial immunodiffusion and immunoelectrophoresis have been discussed under *Proteins*. *IgM* (γM) is the earliest antibody found in the primary immune response and falls rapidly after the onset of IgG antibody synthesis. *IgG* (γG) is the major class of antibody in both the primary and secondary immune response. IgG can cross the placenta to provide the early forms of antibody protection for the newborn. IgG and IgM can participate in the complement fixation reaction. *IgA* (γA) is found predominantly in saliva and secretions of the gastrointestinal and respiratory tracts. In contrast to IgM and IgG, only a small portion of total IgA is found in blood. IgA functions in protection against pathogens that enter the host through the respiratory or gastrointestinal tract. *IgD* (γD) is found in trace quantities in sera and its function is unknown. *IgE* (γE) is probably the most important antibody in acute hypersensitivity or allergic reactions. Reaction of mast cell- or basophil-bound IgE with antigen initiates the release of histamine, slow-reacting substance (SRS), serotonin and bradykinin and the subsequent allergic response. IgE is best quantitated by RIA. Mean serum levels (mg%) in healthy adults are IgG 1200 ± 500, IgA 210 ± 140, IgM 140 ± 70, IgD 3 and IgE <0.1.

Heterophile antibodies are agglutinins which are capable of reacting with antigens that are entirely unrelated to those which stimulate their production. These antibodies, which occur in the serum of patients with infectious mononucleosis or serum sickness, will agglutinate formalized horse erythrocytes. In order to distinguish the specific *heterophile agglutinins of infectious mononucleosis*, the serum sample is mixed with guinea-pig kidney tissue or beef erythrocyte stromata; the infectious mononucleosis antibody will be absorbed and inactivated by the beef cells but not by the kidney tissue, and subsequent agglutination of horse erythrocytes will occur only in the kidney-tissue system. This test is used to detect infectious mononucleosis even prior to clinical symptoms. The heterophile titer has no relation to the course or severity of the disease.

Two protein constituents of human plasma, *rheumatoid factor* (RF) and *C-reactive protein* (CRP) are of value in the differential diagnosis of rheumatoid diseases. CRP is a protein present in the serum of patients in the acute stages of bacterial and viral infections, collagen diseases and other inflammatory processes. The presence of this antigen in serum is detected by agglutination of polystyrene latex particles sensitized with specific CRP antibody globulin. In the management of rheumatic fever, decreases in CRP blood levels are used to measure the effectiveness of therapy.

Rheumatoid arthritis is characterized by the presence of a reactive group of macroglobulins known as RF in blood and synovial fluid. RF is a protein of the IgM globulin fraction and is regarded as an autoantibody against antigenic determinants of IgG. Analysis of RF is based on agglutination procedures employing polystyrene latex particles coated with a layer of adsorbed human gamma globulin. The RF-antibody reaction causes a visible agglutination of the inert latex particles. CRP is not elevated in rheumatoid arthritis.

β-Hemolytic streptococci, the causative agent in rheumatic fever, produce streptolysin O and S, streptokinase, hyaluronidase, desoxyribonuclease and NADase in the body. The growth of streptococci in tissue with elaboration of these proteins serves as the antigenic stimulus to evoke the production of specific antibodies (eg, *antistreptolysin-O, ASO*). The quantitation of the antibody titer to these enzymes is an index of the strength of the antigenic stimulus and the extent of the streptococcal infection. These antibodies can be detected by latex agglutination (ASO) or tests

dependent on the inhibition of enzyme action by the antibody (anti-hyaluronidase inhibition of hyaluronic acid depolymerization by hyaluronidase).

The laboratory diagnosis of *syphilis* (treponemal disease) and the evaluation of a chemotherapeutic approach is based on serological tests. Demonstration of an antibody-like substance, *reagin*, or of true antitreponemal antibody in the serum of infected individuals is accomplished by complement fixation or flocculation tests for reagin, or immunofluorescent techniques for treponemal antibody.

In the *complement fixation* tests (Kolmer CF), reagin reacts with a complex phosphatidic acid antigen (cardiolipin) and complement; the complement is bound and will not lyse hemolysin-sensitized red cells which were added in the second phase of the test. In normal serum the reagin-cardiolipin complex is not formed and the complement is free to react with hemolysin and lyse the erythrocytes.

Flocculation tests for determining of syphilis use a cardiolipin-lecithin-cholesterol antigen which clumps in the presence of serum reagin occurring in nontreponemal diseases and syphilis (*Venereal Disease Research Laboratory—VDRL Test; rapid plasma reagin—RPR test*).

Treponemal antibody can be detected also by the reaction of the patient's serum with treponemal antigen and subsequent confirmation with fluorescein-labeled antihuman globulin as an indicator of primary antigen-antibody reaction (*fluorescent treponemal antibody-FTA test*). The patient's serum can be treated with an extract of treponemes prior to the FTA test to remove interfering antibodies and eliminate biological false-positives (FTA-Abs Test). False-positives occur in related treponematosis such as yaws, pinta and bejel. Increased reagin titers also occur in malaria, leprosy, infectious mononucleosis, chronic rheumatoid arthritis or systemic lupus erythematosus and in patients on hydralazine therapy.

Febrile antibodies are present in the serum of patients with certain bacterial or rickettsial infections (spotted, typhus or Q fever). In typhus the patient's serum contains a febrile antibody which will agglutinate a suspension of *Proteus OX-19* bacteria (Weil-Felix Reaction). *Salmonella* O-H, *Pasteurella tularensis* and *Brucella abortus* antigens are used in febrile antibody tests for diagnosis of typhoid or paratyphoid fever, tularemia and brucellosis, respectively.

Toxoplasmosis is a major cause of birth defects. An expectant mother may become infected with oocysts in uncooked meat, or from cat fur, and infect the fetus transplacentally. Toxoplasmosis testing is based on detecting serum antibody by a hemagglutination procedure. Red cells sensitized by exposure to toxoplasmosis antigen are agglutinated by the specific antibody.

Radioimmunoassay (RIA)[6,27] has been mentioned in various sections of this chapter as an analytical tool in the mensurement of hormones, immunoglobulins, drugs and steroids. The basic principle of RIA is

$$Ag^* + Ag + Ab \leftrightarrows Ag^*Ab + AgAb + Ag^* + Ag \qquad (4)$$

RIA is not to be confused with the *specific reactor assay* using labeled antigen and nonantibody protein receptors which is used for vitamin B_{12}, T^4, T^3 and cortisol assays.

All procedures are based on the observation that radiolabeled antigens (Ag*) compete with nonlabeled antigen (Ag) for binding sites on specific antibody (Ab) in the formation of antigen-antibody complexes (Ag*Ab, AgAb). When increasing amounts of Ag are added to the assay, the binding sites of Ab are saturated progressively and the antibody can bind less Ag*. Therefore, the ratio of bound to free Ag* (B/F) or % Ag* bound is a direct index of the concentration of Ag in the assay.

The requirements for RIA are (1) preparation and characterization of Ag (2) radiolabeling of Ag, (3) preparation of specific Ab and (4) development of the assay system and methods to separate free (Ag, Ag*) from antibody bound (AgAb, Ag*Ab) antigen.

Antigens can be prepared from natural tissue sources or preferably synthesized. ^3H, ^{14}C or ^{125}I-labeled antigens are used routinely in the assay. The biological and immunochemical activity of the antigen must not be altered in the tagging procedure, and the specific activity of Ag* must be extremely high so that tracer quantities can be used in the assay. Tritium labeling and iodination (^{125}I) produce the highest specific activity, but also increase susceptibility of Ag* to internal degradation and self-radiolysis, in contrast to ^{14}C. In many instances, the original antigen cannot be iodinated, but can be altered chemically in such a way as to retain full antigenic cross-reactivity in RIA; eg, cyclic AMP has no tyrosyl or histidyl residue for iodination; ^{125}I-succinylcyclic AMP-tyrosine methyl ester retains full cross-reactivity with antibodies to cyclic AMP and is used in the assay.

Hormones, steroids and drug substances are *haptens*. They do not produce the antibody response when injected by themselves, but will produce antibodies specific for the hapten when injected as a hapten-protein carrier conjugate. Gastrin (hapten) is coupled to albumin (protein-carrier) by treatment with carbodiimides (CCD), which couple functional carboxyl, amino, alcohol, phosphate or thiol groups. Morphine must be converted to the 3-*O*-carboxymethyl derivative prior to CCD coupling with albumin to provide a functional coupling group in the hapten. The hapten-conjugate usually is emulsified in a mineral oil preparation of killed *Mycobacterium* (Complete Freunds Adjuvant) and injected intradermally in rabbits or guinea pigs on several occasions. The serum antibody must have both high specificity and affinity for the antigens.

The *assay system* contains Ag*, sample-containing endogenous Ag or a standard Ag and antibody, at specified pH (6.5 to 8.5). After incubation at 5 to 37° for anywhere from 1 hour to several days, free and antibody-bound antigen must be separated. This is accomplished by *double-antibody technique, solid-phase RIA, resin techniques* or *salt or solvent precipitation*. In the double-antibody technique, antiglobulin (Ab′) serum is added to the assay system after incubation. Ab-Ag* and Ab-Ag complexes are antibody-globulin antigen complexes. The antiglobulin will react to form insoluble Ab′-Ab-Ag* and Ab′-Ab-Ag complexes, which can be removed by centrifugation. The free Ag*, Ag is in the supernate.

The solid phase RIA is performed by coating tubes with Ab; Ag and Ag* react, compete and bind with Ab on the wall of tube. Unreacted Ag and Ag* is separated by decanting and rinsing the tube. Ab also can be bound covalently with isothiocyanate to dextran gel particles. Ag and Ag* will compete and bind with Ab on particles. Bound antigen then can be separated from free antigen by centrifugation.

RIA has been applied to analysis of hormones (ACTH, angiotensin I and II, gastrin, HCG, FSH, GH, glucagon, HLH, HPL, insulin, thyroxine), steroid hormones (aldosterone, androstenedione, glucocorticoids, testosterone, estrones, progesterone), drug substances (digoxin, digitoxin, amphetamines, barbiturates, morphine, LSD, ouabain), endogenous substances (cyclic AMP, cyclic GMP, prostaglandins, immunoglobulins, hepatitis antigen, carcinoembryonic antigen—CEA). Examples of the specific assays are discussed in other sections.

CEA and AFP (α-1-fetoprotein) are proteins found in fetal tissue. CEA analysis was first proposed as a specific test for the early detection of bowel cancer. Although the test does not have absolute specificity for this disease, it may prove of value as a diagnostic aid and therapy monitor. CEA can be detected by RIA. Serum levels >2.5 ng CEA/mL are found in 60 to 70% of patients with adenocarci-

noma of the colon; positive levels also are found in lower percentages in carcinomas of the pancreas, stomach, liver, breast, endometrium, ovary, kidney and bronchus, as well as in other conditions such as gastrointestinal polyps, colitis, diverticulitis and cirrhosis. CEA appears to be associated primarily with tumors of entodermally derived epithelial tissue. The similarity between CEA and cell-surface glycoproteins and sialic acids has stimulated considerable research interest in a new approach to cancer chemotherapy.

The study of *tissue-transplantation antigens* is an important factor in studies on tissue and organ transplants. ABO blood group antigens are involved in survival of skin and renal grafts. Because of the presence of natural occurring anti- A and B, avoidance of ABO incompatibility is important in clinical grafting. The *HL-A antigens* are found on tissue and on the white cells. There is one major histocompatibility locus, comprising a number of alleles or linked genes, on a single chromosome segment. Each allele controls four to five groups of major transplantation antigens. These HL-A isoantigens affect the survival of allogenic tissue grafts and organ transplants. HL-A antigens can be typed by a leukoagglutination method in which the patient's or donor's white cells are reacted with specific HL-A antisera. HL-A typing also can be performed by a cytotoxicity test in which lymphocytes are mixed with antisera and complement. The antibody can destroy the lymphocytes if a corresponding antigen is present on the cell surface.

References

1. Mitruka BM, Rawnsley AM: *Clinical and Hematological Reference Values in Normal Experimental Animals and Normal Humans*, Yearbook Medical Pub, Chicago, 1981.
2. Christensen RL, Triplett DA: *Lab Med, 13(11):* 666-672, 1982.
3. Bollinger P, Brailas CD, Drewinko B: Evaluation of whole-blood platelet analyzers. *Ibid 14:* 492, 1983.
4. *Central File for Rare Donors*, Am Assoc Blood Banks, Milwaukee.
5. *ABO and Rh Systems*, Ortho Diagnostics, Raritan NJ 1969.
6. Berson S, Yalow R: *Gastroenterol 62:* 1061, 1972.
7. *Fed Reg 37FR17419*, Aug 26, 1972.
8. Broughton PMG, Dawson JB: *Advan Clin Chem 15:* 288, 1972.
9. Mears T, Young D: *Am J Clin Pathol 50:* 411, 1968.
10. Meinke W: *Anal Chem 43:* 28A, 1971.
11. Dybaker R: *Std Methods Clin Chem 6:* 223, 1970.
12. Flokstra JH, Soda JA: *Am J Med Sci 254:* 429, 1967.
13. Young DS *et al: Clin Chem 21:* 1D, 1975.
14. Constantino NV, Kabat HF: *Am J Hosp Pharm 30:* 24, 1973.
15. Peterson CM: *Diagn Med:* 78, Jul/Aug 1980.
16. *Radial Immunodiffusion and Immunoelectrophoreses for Qualitation and Quantitation of Immunoglobulins* (DHEW Publ HSM-72-8102), USD-HEW, Washington DC, 1972.
17. *Bull WHO 43:* 891, 1970.
18. Statland BE: *Clinical Decision Levels For Lab Tests*, Med Econ Co Inc., Oradell NJ, 1983.
19. White W *et al: Practical Automation for the Clinical Laboratory*, Mosby, St Louis, 1972.
20. Szilagyi G, Aning V, Karmen A: *J Clin Lab Automation 3:* 117, 1983.
21. Bradley GM: Fecal analysis. *Diagn Med* 63 Mar/Apr 1980.
22. Rubenstein K *et al: Biochem Biophys Res Comm 47:* 846, 1972.
23. Baron J: *Scand J Gastroenterol 5:* 9, 1970.
24. Jorpes J, Mott V: *Secretin, CCK, Pancreozymin and Gastrin*, Springer Verlag, New York, 1973.
25. Kuhn PJ: *Med Lab Observer* 108, Sept 1983.
26. Feldman M, Nossal GJ: *Quart Rev Biol 47:* 269, 1972.
27. Skelley DS, *et al: Clin Chem 19:* 146, 1973.

Bibliography

Wintrobe M: *Clinical Hematology*, 6th ed, Lea & Febiger, Philadelphia, 1967.

Lynch MJ: *Medical Laboratory Technology*, 2nd ed, Saunders, Philadelphia, 1969.

Faulkner W, King J: *Manual Clinical Laboratory Procedures*, Chemical Rubber Co, Cleveland, 1970.

Faulkner W *et al: Handbook Clinical Laboratory Data*, Chemical Rubber Co, Cleveland, 1980.

Roth K, Saunders A: *Evaluation of Methods for White Cell Identification and Counting-Advances in Automated Analysis*, Technicon Intl Congr, 1970.

Dacie J, Lewis S: *Practical Hematology*, 3rd ed, Churchill, London, 1963.

Frankel S, Reitman S, eds: *Clinical Laboratory Methods and Diagnosis*, 6th ed, Mosby, St. Louis, 1963.

Wintrobe MM: *Laboratory Medicine-Hematology*, 2nd ed, Mosby, St. Louis, 1962.

Manual of Blood Coagulation Technics, 2nd ed, Warner-Chilcott, Morris Plans NJ, 1966.

Detection of Fibrinogen Degradation Products, Wellcome Res Labs, England, 1973.

A Manual of Methods for the Coagulation Laboratory, BD & Co, Rutherford NJ, 1965.

Technical Methods and Procedures of the American Association of Blood Banks, Am Assoc Blood Banks, Chicago, 1962.

Standards for Blood Transfusion Service, 4th ed, Am Assoc Blood Banks, Chicago, 1963.

Griffiths JJ, Elliott J: *Blood Bank Procedures*, Dade Reagents, Miami, 1967.

Chromatography in Mass Screening for Disorders of Amino Acid Metabolism, Hyland, Los Angeles, 1966.

Rosalki S, Wilkinson J: *Diagnostic Enzymology*, Dade Reagents, Miami, 1966.

Wilkinson J: *Introduction to Diagnostic Enzymology*, Edward Arnold, Ltd. London, 1962.

Davidsohn I, Henry J: *Todd-Sanford Clinical Diagnosis by Laboratory Methods*, 15th ed, Saunders, Philadelphia, 1974.

Peron FG, Caldwell BV: *Immunologic Methods in Steroid Determination*, Appleton-Century, New York, 1970.

Winsten S, Dalal F: *Clinical Laboratory Procedures for Nonroutine Problems*, Chemical Rubber Co, Cleveland, 1972.

Specialized Diagnostic Laboratory Tests, Bioscience Labs, Van Nuys CA, 1971.

Kark RM *et al: A Primer of Urinalysis*, 2nd ed, Harper & Row, New York, 1963.

Sunderman FW, Sunderman FW, Jr: *Laboratory Diagnosis of Renal Diseases*, Warren H Green, Inc, St Louis, 1970.

Faust E, Russell P: *Clinical Parasitology*, 7th ed, Lea & Febiger, Philadelphia, 1964.

Sunshine I: *Manual of Analytical Toxicology*, Chemical Rubber Co., Cleveland, 1972.

Clarke, E: *Isolation and Identification of Drugs*, vols 1 and 2, Pharmaceutical Press, London, 1969 and 1975.

Blair JE *et al: Manual of Clinical Microbiology*, Williams & Wilkins, Baltimore, 1970.

Edwards PR, Ewing WH: *Identification of Enterobacteriaceae*, 3rd ed, Burgess, Minneapolis, 1972.

Holdeman LV, Moore WEC: *Anaerobe Laboratory Manual*, 2nd ed, VPI Anaerobe Lab, Blacksburg VA, 1973.

Connant NF *et al: Manual of Clinical Mycology*, 3rd ed, Saunders, Philadelphia, 1971.

Buch F, Good R: *Clinical Immunobiology*, Academic, New York, 1972.

Clinical RIA. *Lab Manag:* May 1973.

Manual of tissue typing techniques. *Natl Inst All Infect Dis Bull:* 1972.

Directory of Rare Analysis. *Clin Chem 23:* 323, 1977.

Doucet LD, *Medical Technology Review*, Lippincott, Philadelphia, 1981.

Hansten PD, *Drug Interactions*, 3rd ed, Lea & Febiger, Philadelphia, 1975.

Kaplan A, Szabo LL, *Clinical Chemistry: Interpretation and Techniques*, 2nd ed, Lea & Febiger, Philadelphia, 1983.

Miller SE, Weller JM, *Textbook of Clinical Pathology*, 8th ed, Williams & Wilkins, Baltimore, 1971.

Pencock J, Tomar R, *Manual of Laboratory Immunology*, Lea & Febiger, Philadelphia, 1980.

Pertinent Reference Journals

Advan Clin Chem	*J Clin Lab Automation*
Am J Clin Pathol	*Diagn Med*
Am Clin Prod Rev	*J Lab Clin Med*
Am J Hosp Pharm	*Lab Med*
Am J Med Technol	*Lab Notes Med Diag*
Anal Chem	*Med Lab Obs*
BioTechniques	*Med Lab Tech*
Clin Chem	*Scand J Clin Lab Invest*
Clin Chim Acta	*Std Methods Clin Chem*

CHAPTER 35

Drug Absorption, Action and Disposition

Stewart C Harvey, PhD

Professor of Pharmacology
School of Medicine, University of Utah
Salt Lake City, UT 84132

Although drugs differ widely in their pharmacodynamic effects and clinical application, in penetrance, absorption and usual route of administration, in distribution among the body tissues and in disposition and mode of termination of action, there are certain general principles that help explain these differences. These principles have both pharmaceutic and therapeutic implications. They facilitate an understanding of both the features that are common to a class of drugs and the differentia among the members of that class.

In order for a drug to act it must be absorbed, transported to the appropriate tissue or organ, penetrate to the responding subcellular structure and elicit a response or alter ongoing processes. The drug may be distributed simultaneously or sequentially to a number of tissues, bound or stored, metabolized to inactive or active products or excreted. The history of a drug in the body is summarized in Fig 35-1. Each of the processes or events depicted relates importantly to therapeutic and toxic effects of a drug and to the mode of administration, and drug design must take each into account. Since the effect elicited hy a drug is its *raison d'etre*, *drug action* and *effect* will be discussed first in the text that follows, even though they are preceded by other events.

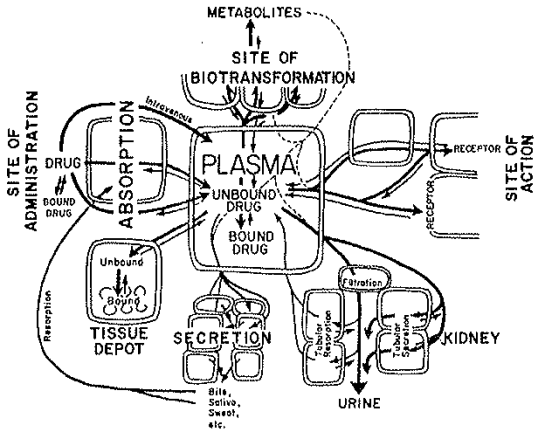


Fig 35-1. The absorption, distribution, action and elimination of a drug (arrows represent drug movement). Intravenous administration is the only process whereby a drug may enter a compartment without passing through a biological membrane. Note that drugs excreted in bile and saliva may be resorbed.

Drug Action and Effect

The word *drug* imposes an action–effect context within which the properties of a substance are described. The description of necessity must include the pertinent properties of the recipient of the drug. Thus, when a drug is defined as an analgesic, it is implied that the recipient reacts in a certain way, called pain,* to a noxious stimulus. Both because the pertinent properties are locked into the complex and somewhat imprecise biological context and because the types of possible response are many, descriptions of the properties of drugs tend to emphasize the qualitative features of the effects they elicit. Thus, a drug may be described as having analgesic, vasodepressor, convulsant, antibacterial, etc, properties. The specific effect (or use) categories into which the many drugs may be placed are the subject of Chapters 38 through 65 and will not be elaborated upon in this chapter. However, the description of a drug does not end with the enumeration of the responses it may elicit. There are certain intrinsic properties of the drug–recipient system that can be described in quantitative terms and which are essential to the full description of the drug and to the validation of the drug for specific uses. Under *Definitions and Concepts*, below, certain general terms are defined in qualitative language; under *Dose–Effect Relationships* the foundation is laid for an appreciation of some of the quantitative aspects of pharmacodynamics.

Definitions and Concepts

In the field of pharmacology, the vocabulary that is unique to the discipline is relatively small, and the general vocabulary is that of the biological sciences and chemistry. Nevertheless, there are a few definitions that are important to the proper understanding of pharmacology. It is necessary to differentiate among action, effect, selectivity, dose, potency and efficacy.

Action vs Effect—The *effect* of a drug is an *alteration of function* of the structure or process upon which the drug acts. It is common to use the term action as a synonym for effect. However, action precedes effect. *Action* is the *alteration of condition* that brings about the effect.

The final effect of a drug may be far removed from its site of action. For example, the diuresis subsequent to the ingestion of ethanol does not result from an action on the kidney but instead from a depression of activity in the supraopticohypophyseal region of the hypothalamus, which regulates the release of antidiuretic hormone from the posterior pituitary gland. The alteration of supraopticohypo-

* Sophisticated studies indicate that pain is not simply the *perception* of a certain kind of stimulus but, rather, a *reaction* to the perception of a variety of kinds of stimuli or stimulus patterns.

physeal function is, of course, also an effect of the drug, as is each subsequent change in the chain of events leading to diuresis. The action of ethanol was exerted only at the initial step, each subsequent effect being then the action to a following step.

Multiple Effects—No known drug is capable of exerting a single effect, although a number are known that appear to have a single mechanism of action. Multiple effects may derive from a single mechanism of action. For example, the inhibition of acetylcholinesterase by physostigmine will elicit an effect at every site where acetylcholine is produced, is potentially active, and is hydrolyzed by cholinesterase. Thus, physostigmine elicits a constellation of effects.

A drug also can cause multiple effects at several different sites by a single action at only one site, providing that the function initially altered at the site of action ramifies to control other functions at distant sites. Thus, a drug that suppresses steroid synthesis in the liver may not only lower serum cholesterol, impair nerve myelination and function and alter the condition of the skin (as a consequence of cholesterol deficiency) but also may affect digestive functions (because of a deficiency in bile acids) and alter adrenocortical and sexual hormonal balance.

Although a single action can give rise to multiple effects, most drugs exert multiple actions. The various actions may be related, as, for example, the sympathomimetic effects of metaraminol that accrue to its structural similarity to norepinephrine and its ability partially to suppress sympathetic responses because it occupies the catecholamine storage pools in lieu of norepinephrine; or the actions may be unrelated, as with the actions of morphine to interfere with the release of acetylcholine from certain autonomic nerves, block some actions of 5-hydroxytryptamine (serotonin) and release histamine. Many drugs bring about immunologic (allergic or hypersensitivity) responses that bear no relation to the other pharmacodynamic actions of the drug.

Selectivity—Despite the potential most drugs have for eliciting multiple effects, one effect is generally more readily elicitable than another. This differential responsiveness is called *selectivity*. It usually is considered to be a property of the drug, but it is also a property of the constitution and biodynamics of the recipient subject or patient.

Selectivity may come about in several ways. The subcellular structure (receptor) with which a drug combines to initiate one response may have a higher affinity for the drug than that for some other action. Atropine, for example, has a much higher affinity for muscarinic receptors (page 889) that subserve the function of sweating than it does for the nicotinic receptors (page 889) that subserve voluntary neuromuscular transmission, so that suppression of sweating can be achieved with only a tiny fraction of the dose necessary to cause paralysis of the skeletal muscles. A drug may be distributed unevenly, so that it reaches a higher concentration at one site than throughout the tissues generally; chloroquine is much more effective against hepatic than intestinal (colonic) amebiasis because it reaches a much higher concentration in the liver than in the wall of the colon. An affected function may be much more critical to or have less reserve in one organ than in another, so that a drug will be predisposed to elicit an effect at the more critical site. Some inhibitors of dopa decarboxylase (which is also 5-hydroxytryptophan decarboxylase) depress the synthesis of histamine more than that of either norepinephrine or 5-hydroxytryptamine (serotonin), even though histidine decarboxylase is less sensitive to the drug, simply because histidine decarboxylase is the only step and, hence, is rate-limiting in the biosynthesis of histamine. Dopa decarboxylase is not rate-limiting in the synthesis of either norepinephrine or 5-hydroxytryptamine until the enzyme is nearly completely inhibited. Another example of the determination of selectivity by the critical balance of the affected function is that of the mercurial diuretic drugs. An inhibition of only 1% in the tubular resorption of glomerular filtrate usually will double urine flow, since 99% of the glomerular filtrate is normally resorbed. Aside from the question of the possible concentration of diuretics in the urine, a drug-induced reduction of 1% in sulfhydryl enzyme activity in tissues other than the kidney usually is not accompanied by an observable change in function. Selectivity also can be determined by the pattern of distribution of destructive or activating enzymes among the tissues and by other factors.

Dose—Even the uninitiated person knows that the *dose* of a drug is the amount administered. However, the appropriate dose of a drug is not some unvarying quantity, a fact sometimes overlooked by pharmacists, official committees and physicians. The practice of pharmacy is entrapped in a system of fixed-dose formulations, so that fine adjustments in dosage are often difficult to achieve. Fortunately, there is usually a rather wide latitude allowable in dosages. It is obvious that the size of the recipient individual should have a bearing upon the dose, and the physician may elect to administer the drug on a body-weight or surface-area basis rather than as a fixed dose. Usually, however, a fixed dose is given to all adults, unless the adult is exceptionally large or small. The dose for infants and children often is determined by one of several formulas which take into account age or weight, depending on the age group of the child and the type of action exerted by the drug. Infants, relatively, are more sensitive to many drugs, often because enzyme systems which destroy the drugs may not be developed fully in the infant.

The nutritional condition of the patient, the mental outlook, the presence of pain or discomfort, the severity of the condition being treated, the presence of secondary disease or pathology, genetic and many other factors affect the dose of a drug necessary to achieve a given therapeutic response or to cause an untoward effect (Chapter 67). Even two apparently well-matched normal persons may require widely different doses for the same intensity of effect. Furthermore, a drug is not always employed for the same effect and, hence, not in the same dose. For example, the dose of a progestin necessary for an oral contraceptive effect is considerably different from that necessary to prevent spontaneous abortion, and a dose of an estrogen for the treatment of the menopause is much too small for the treatment of prostatic carcinoma.

From the above it is evident that the wise physician knows that *the dose of a drug is "enough"* (ie, no rigid quantity but rather that which is necessary and can be tolerated) and individualizes the regimen accordingly. The wise pharmacist also will appreciate this dictum and recognize that official or manufacturer's recommended doses are sometimes quite narrowly defined and may be very wide of the mark. They should serve only as a useful guide rather than as an imperative.

Potency and Efficacy—The *potency* of a drug is the reciprocal of dose. Thus, it will have the units of persons/unit weight of drug or body weight/unit weight of drug, etc. Potency generally has little utility other than to provide a means of comparing the relative activities of drugs in a series, in which case *relative potency*, relative to some prototype member of the series, is a parameter commonly used among pharmacologists and in the pharmaceutical industry.

Whether a given drug is more potent than another has little bearing on its clinical usefulness, provided that the potency is not so low that the size of the dose is physically unmanageable or the cost of treatment is higher than with an equivalent drug. If a drug is less potent but more selective, it is the one to be preferred. Promotional arguments in favor of a more potent drug thus are irrelevant to the impor-

tant considerations that should govern the choice of a drug. However, it sometimes occurs that drugs of the same class differ in the maximum intensity of effect; that is, some drugs of the class may be less efficacious than others, irrespective of how large a dose is used.

Efficacy connotes the property of a drug to achieve the desired response, and *maximum efficacy* denotes the maximum achievable effect. Even huge doses of codeine often cannot achieve the relief from severe pain that relatively small doses of morphine can; thus, codeine is said to have a lower maximum efficacy than morphine. Efficacy is one of the primary determinants of the choice of a drug.

Dose–Effect Relationships

The importance of knowing how changes in the intensity of response to a drug vary with the dose is virtually self-evident. Both the physician, who prescribes or administers a drug, and the manufacturer, who must package the drug in appropriate dose sizes, must translate such knowledge into everyday practice. Theoretical or molecular pharmacologists also study such relationships in inquiries into mechanism of action and receptor theory (see page 702). It is necessary to define two types of relationships: (1) dose–intensity relationship—ie, the manner in which the intensity of effect in the individual recipient relates to dose—and (2) dose–frequency relationship—ie, the manner in which the number of responders among a population of recipients relates to dose.

Dose–Intensity of Effect Relationships—Whether the intensity of effect is determined *in vivo* (eg, the blood-pressure response to epinephrine in the human patient) or *in vitro* (eg, the response of the isolated guinea pig ileum to histamine), the dose–intensity of effect (often called dose-effect) curve usually has a characteristic shape, namely a curve that closely resembles one quadrant of a rectangular hyperbola.

In the dose–intensity curve depicted in Fig 35-2, the curve appears to intercept the x axis at 0 only because the lower doses are quite small on the scale of the abscissa, the smallest dose being 1.5×10^{-3} µg. Actually, the x intercept has a positive value, since a finite dose of drug is required to bring about a response, this lowest effective dose being known as the *threshold dose*. Statistics and chemical kinetics predict

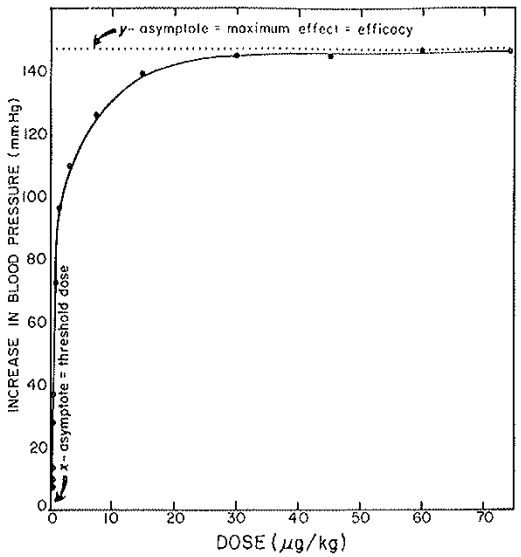


Fig 35-2. The relationship of the intensity of the blood-pressure response of the cat to the intravenous dose of norepinephrine.

that the curve should approach the y axis asymptotically. However, if the intensity of the measured variable does not start from zero, the curve possibly may have a positive y intercept (or negative x intercept), especially if the ongoing basal activity before the drug is given is closely related to that induced by the drug.

In practice, instead of an asymptote to the y axis, dose–intensity curves nearly always show an upward concave foot at the origin of the curve, so that the curve has a lopsided sigmoid shape. At high doses, the curve approaches an asymptote which is parallel to the x axis, and the value of the asymptote establishes the maximum possible response to the drug, or *maximum efficacy*. However, experimental data in the regions of the asymptotes generally are too erratic to permit an exact definition of the curve at the very low and very high doses. The example shown represents an unusually good set of data.

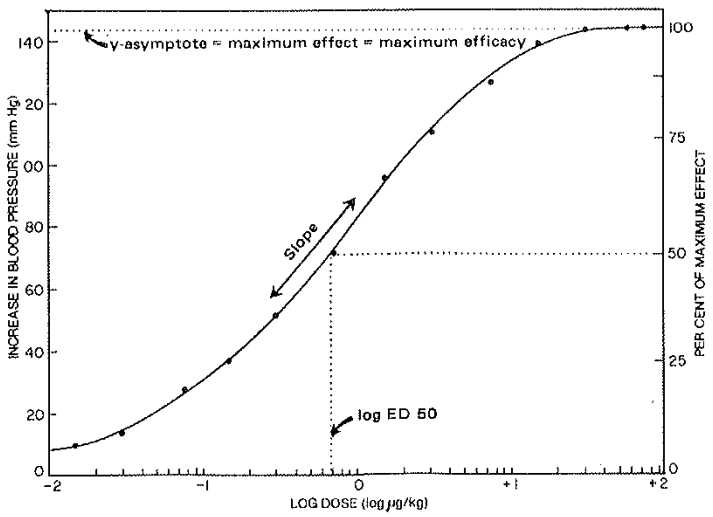


Fig 35-3. The relationship of the intensity of the blood-pressure response of the cat to the log of the intravenous dose of norepinephrine.

Because the dose range may be 100- or 1000-fold from the lowest to the highest dose, it has become the practice to plot dose–intensity curves on a logarithmic scale of abscissa; ie, to plot the log of dose versus the intensity of effect. Figure 35-3 is such a semilogarithmic plot of the same data as in Fig 35-2. In the figure the intensity of effect is plotted both in absolute units (at the left) or in relative units, as percent (at the right).

Although no new information is created by a semilogarithmic representation, the curve is stretched in such a way as to facilitate the inspection of the data; the comparison of results from multiple observations and the testing of different drugs also is rendered easier. In the example shown, the curve is essentially what is called a *sigmoid curve* and is nearly symmetrical about the point which represents an intensity equal to 50% of the maximal effect, ie, about the midpoint. The symmetry follows from the rectangular hyperbolic character of the previous Cartesian plot (Fig 35-2). The semilogarithmic plot reveals better the dose–effect relationships in the low-dose range, which are lost in the steep slope of the Cartesian plot. Furthermore, the data about the midpoint are almost a straight line; the nearly linear portion covers approximately 50% of the curve. The slope of the linear portion of the curve or, more correctly, the slope at the point of inflection, has theoretical significance (see *Drug Receptors and Receptor Theory*, page 702).

The upper portion of the curve approaches an asymptote, which is the same as that in the Cartesian plot. If the response system is completely at rest before the drug is administered, the lower portion of the curve should be asymptotic to the x axis. Both asymptotes and the symmetry derive from the law of mass action (see page 703).

Dose–intensity curves often deviate from the ideal configuration illustrated and discussed above. Usually, the deviate curve remains sigmoid but not extended symmetrically about the midpoint of the *linear* segment. Occasionally, other shapes occur, sometimes quite bizarre ones. Deviations may derive from multiple actions that converge upon the same final effector system, from varying degrees of metabolic alteration of the drug at different doses, from modulation of the response by feedback systems, from nonlinearity in the relationship between action and effect or from other causes.

It is frequently necessary to identify the dose which elicits a given intensity of effect. The intensity of effect that is generally designated is the 50% of maximum intensity. The corresponding dose is called the *50% effective dose*, or *individual ED50* (see Fig 35-3). The use of the adjective, *individual*, distinguishes the ED50 based upon the intensity of effect from the median effective dose, also abbreviated ED50, determined from frequency of response data in a population (see *Dose–Frequency Relationships*, this page).

Drugs that elicit the same quality of effect may be compared graphically. In Fig 35-4, five hypothetical drugs are compared. Drugs *A*, *B*, *C* and *E* all can achieve the same maximum effect, which suggests that the same effector system may be common to all. *D* possibly may be working through the same effector system, but there are no *a priori* reasons to think this is so. Only *A* and *B* have parallel curves and common slopes. Common slopes are consistent with, but in no way prove, the idea that *A* and *B* not only act through the same effector system but also by the same mechanism. Although drug–receptor theory (see *Drug Receptors and Receptor Theory*, page 702) requires that the curves of identical mechanism have equal slopes, examples of exceptions are known. Furthermore, mass-law statistics require that all simple drug-receptor interactions generate the same slope; only when slopes depart from this universal slope in accordance with distinctive characteristics of the response system do they provide evidence of specific mechanisms.

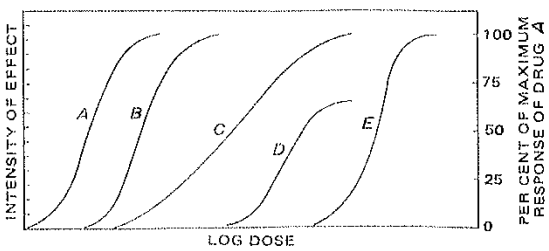


Fig 35-4. Log dose–intensity of effect curves of five different hypothetical drugs (see text for explanation).

The relative potency of any drug may be obtained by dividing the ED50 of the standard, or prototype, drug by that of the drug in question. Any level of effect other than 50% may be used, but it should be recognized that when the slopes are not parallel the relative potency depends upon the intensity of effect chosen. Thus, the potency of *A* relative to *C* (in Fig 35-4) calculated from the ED50 will be smaller than that calculated from the ED25.

The low maximum intensity inducible by *D* poses even more complications in the determination of relative potency than do the unequal slopes of the other drugs. If its dose–intensity curve is plotted in terms of percent of its own maximum effect, its relative inefficacy is obscured and the limitations of relative potency at the ED50 level will not be evident. This dilemma simply underscores the fact that drugs can be compared better from their entire dose–intensity curves than from a single derived number like ED50 or relative potency.

Drugs that elicit multiple effects will generate a dose–intensity curve for each effect. Even though the various effects may be qualitatively different, the several curves may be plotted together on a common scale of abscissa, and the intensity may be expressed in terms of percent of maximum effect; thus, all curves can share a common scale of ordinates in addition to common abscissa. Separate scales of ordinates could be employed, but this would make it harder to compare data.

The selectivity of a drug can be determined by noting what percent of maximum of one effect can be achieved before a second effect occurs. As with relative potency, selectivity may be expressed in terms of the ratio between the ED50 for one effect to that for another effect, or a ratio at some other intensity of effect. Similarly to relative potency, difficulties follow from nonparallelism. In such instances, selectivity expressed in dose ratios varies from one intensity level to another.

When the dose–intensity curves for a number of subjects are compared, it is found that they vary considerably from individual to individual in many respects; eg, threshold dose, midpoint, maximum intensity and sometimes even slope. By averaging the intensities of the effect at each dose, an average dose–intensity curve can be constructed.

Average dose–intensity curves enjoy a limited application in comparing drugs. A single line expressing an average response has little value in predicting individual responses unless it is accompanied by some expression of the range of the effect at the various doses. This may be done by indicating the standard error of the response at each dose. Occasionally, a simple scatter diagram is plotted in lieu of an average curve and statistical parameters (see Fig 10-21). An average dose–intensity curve also may be constructed from a population in which different individuals receive different doses; if sufficiently large populations are employed, the average curves determined by the two methods will approximate each other.

It is obvious that the determination of such average curves from a population sufficiently large to be statistically meaningful requires a great deal of work. Retrospective clinical data occasionally are treated in this way, but prospective studies infrequently are designed in advance to yield average curves. The usual practice in comparing drugs is to employ a quantal (all-or-none) end-point and plot the frequency or cumulative frequency of response over the dose range, as discussed below.

Dose–Frequency of Response Relationships—When an end-point is truly all-or-none, such as death, it is an easy matter to plot the number of responding individuals (eg, dead subjects) at each dose of drug or intoxicant. Many other responses that vary in intensity can be treated as all-or-none if simply the presence or absence of a response (eg, cough or no cough, convulsion or no convulsion) is recorded, without regard to the intensity of the response when it occurs.

When the response changes from the basal or control state in a less abrupt manner (eg, tachycardia, miosis, rate of gastric secretion) it may be necessary to designate arbitrarily some particular intensity of effect as the end-point. If the end-point is taken as an increase in heart rate of 20 beats/min, all individuals whose tachycardia is less than 20/min would be recorded as nonresponders, while all those with 20 or above would be recorded as responders. When the percent of responders in the population is plotted against the dose, a characteristic dose–response curve, more properly called a *dose–cumulative frequency* or *dose–percent* curve, is generated. Such a curve is, in fact, a cumulative frequency–distribution curve, the percent of responders at a given dose being the frequency of response.

Dose–cumulative frequency curves are generally of the same geometric shape as dose–intensity curves (namely, sigmoid) when frequency is plotted against log dose (see Fig 35-5). The tendency of the cumulated frequency of response (ie, percent) to be linearly proportional to the log of the dose in the middle of the dose range is called the *Weber–Fechner law*, although it is not invariable, as a true natural law should be. In many instances, the cumulative frequency is simply proportional to dose rather than log dose. The Weber–Fechner law applies to either dose–intensity or dose–cumulative frequency data. The similarity between dose–frequency and dose–intensity curves may be more than fortuitous, since the intensity of response will usually have an approximately linear relationship to the percent of responding *units* (smooth muscle cells, nerve fibers, etc) and, hence, is also a type of cumulative frequency of response. These are the same kind of statistics that govern the law of mass action.

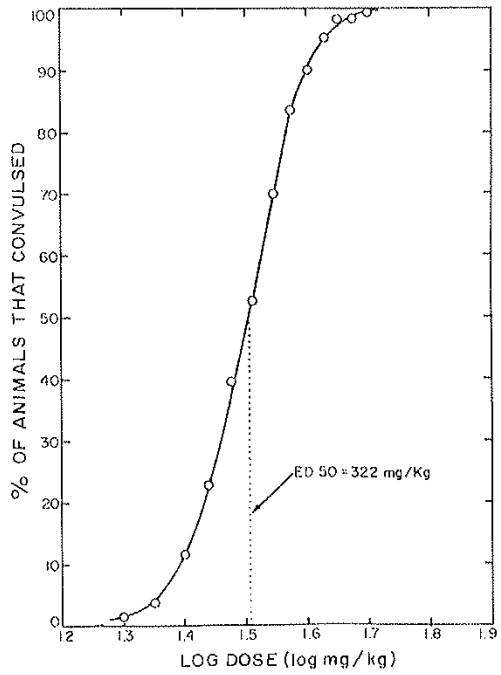


Fig 35-5. The relationship of the number of responders in a population of mice to the dose of pentylenetetrazol (courtesy, Drs DG McQuarry and EG Fingl, University of Utah).

If only the increase in the number of responders with each new dose is plotted, instead of the cumulative percent of responders, a bell-shaped curve is obtained. This curve is the first derivative of the dose–cumulative frequency curve and is a *frequency–distribution* curve (see Chapter 10). The distribution will be symmetrical—ie, *normal* or Gaussian (see Fig 10-5)—only if the dose–cumulative frequency curve is symmetrically hyperbolic. Because most dose–cumulative frequency curves are more nearly symmetrical when plotted semilogarithmically (ie, as log dose), dose–cumulative frequency curves are usually *log-normal*.

Since the dose–intensity and dose–cumulative frequency curves are basically similar in shape, it follows that the curves have similar defining characteristics, such as ED50, maximum effect (maximum efficacy) and slope. In dose–cumulative frequency data, the ED50 (*median effective dose*) is the dose to which 50% of the population responds (see Fig 35-5). If the frequency distribution is normal, the ED50 is both the arithmetic mean and median dose and is represented by the midpoint on the curve; if the distribution is log-normal, the ED50 is the median dose but not the arithmetic mean dose. The efficacy is the cumulative frequency summed over all doses; it is usually, but not always, 100%. The slope is characteristic of both the drug and test population. Even two drugs of identical mechanism may give rise to different slopes in dose–percent curves, whereas in dose–intensity curves the slopes are the same.

Statistical parameters (such as standard deviation), in addition to ED50, maximum cumulative frequency (efficacy) and slope, characterize dose-cumulative frequency relationships (see Chapter 10).

There are several formulations for dose–cumulative frequency curves, some of which are employed only to define the linear segment of a curve and to determine the statistical parameters of this segment. For the statistical treatment of dose–frequency data, see Chapter 10. One simple mathematical expression of the entire log-symmetrical sigmoid curve is

$$\log \text{dose} = K + f \log\left(\frac{\%\ \text{response}}{100\% - \text{response}}\right) \qquad (1)$$

where percent response may be either the percent of maximum intensity or the percent of a population responding. The equation is thus basically the same for both log normal dose–intensity and log normal dose–percent relationships. K is a constant that is characteristic of the midpoint of the curve, or ED50, and $1/f$ is characteristically related to the slope of the linear segment, which, in turn is closely related to the standard deviation of the derivative log-normal frequency-distribution curve.

The comparison of dose–percent relationships among drugs is subject to the pitfalls indicated for dose–intensity comparisons (see page 699), namely, that when the slopes of the curves are not the same (ie, the dose–percent curves are not parallel), it is necessary to state at which level of response a potency ratio is calculated. As with dose–intensity

data, potencies generally are calculated from the ED50, but potency ratios may be calculated for any arbitrary percent response. The expression of selectivity is, likewise, subject to similar qualifications, inasmuch as the dose–percent curves for the several effects are usually nonparallel.

The term *therapeutic index* is used to designate a quantitative statement of the selectivity of a drug when a therapeutic and an untoward effect are being compared. If the untoward effect is designated as *T* (for toxic) and the therapeutic effect as *E*, the therapeutic index may be defined as TD50/ED50 or a similar ratio at some other arbitrary levels of response. The TD and the ED are not required to express the same percent of response; some clinicians use the ratio TD1/ED99 or TD5/ED95, based on the rationale that if the untoward effect is serious, it is important to use a most-severe therapeutic index in passing judgment upon the drug. Unfortunately, therapeutic indices are known in man for only a few drugs.

There will be a different therapeutic index for each untoward effect that a drug may elicit, and, if there is more than one therapeutic effect, a family of therapeutic indices for each therapeutic effect. However, in clinical practice, it is customary to distinguish among the various toxicities by indicating the percent incidence of a given side effect.

Variations in Response and Responsiveness—From the above discussion of dose–frequency relationships and Chapter 10, it is obvious that in a normal population of persons there may be quite a large difference in the dose required to elicit a given response in the least-responsive member of the population and that to elicit the response in the most-responsive member. The difference ordinarily will be a function of the slope of the dose–percent curve, or, in statistical terms, of the standard deviation. If the standard deviation is large, the extremes of responsiveness of responders are likewise large.

In a normal population 95.46% of the population responds to doses within two standard deviations from the ED50 and 99.73% within three standard deviations. In log-normal populations the same distribution applies when standard deviation is expressed as log dose.

In the population represented in Fig 35-5, 2.25% of the population (two standard deviations from the median) would require a dose more than 1.4 times the ED50; an equally small percent would respond to 0.7 the ED50. The physician who is unfamiliar with statistics is apt to consider the 2.25% at either extreme as abnormal reactors. The statistician will argue that these 4.5% are within the normal population and only those who respond well outside of the normal population, at least three standard deviations from the median, deserve to be called abnormal.

Irrespective of whether the criteria of abnormality that the physician or the statistician obtain, the term *hyporeactive* applies to those individuals who require abnormally high doses and *hyperreactive* to those who require abnormally low doses. The terms *hyporesponsive* and *hyperresponsive* also may be used. It is incorrect to use the terms hyposensitive and hypersensitive in this context; *hypersensitivity* denotes an allergic response to a drug and should not be used to refer to hyperreactivity. The term *supersensitivity* correctly applies to hyperreactivity that results from denervation of the effector organ; it is often more definitively called denervation supersensitivity. Sometimes hyporeactivity is the result of an immunochemical deactivation of the drug, or *immunity*. Hyporeactivity should be distinguished from an increased dose requirement that results from a severe pathological condition. Severe pain requires large doses of analgesics, but the patient is not a hyporeactor; what has changed is the baseline from which the endpoint quantum is measured. The responsiveness of a patient to certain drugs sometimes may be determined by the history of previous exposure to appropriate drugs.

Tolerance is a diminution in responsiveness as use of the drug continues. The consequence of tolerance is an increase in the dose requirement. It may be due to an increase in the rate of elimination of drug (as discussed elsewhere in this chapter), to reflex or other compensatory homeostatic adjustments, to a decrease in the number of receptors or in the number of enzyme molecules or other coupling proteins in the effector sequence, to exhaustion of the effector system or depletion of mediators, to the development of immunity or to other mechanisms. Tolerance may be gradual, requiring many doses and days to months to develop, or acute, requiring only the first or a few doses and only minutes to hours to develop. Acute tolerance is called *tachyphylaxis*.

Drug resistance is the decrease in responsiveness of microorganisms, neoplasms or pests to chemotherapeutic agents, antineoplastics or pesticides, respectively. It is not tolerance in the sense that the sensitivity of the individual microorganism or cancer cell decreases; rather, it is the survival of normally unresponsive cells which then pass the genetic factors of resistance on to their progeny.

Patients who fail to respond to a drug are called *refractory*. Refractoriness may result from tolerance or resistance, but it also may result from the progression of pathological states that negate the response or render the response incapable of surmounting an overwhelming pathology. Rarely, it may result from a poorly developed receptor or response system.

Sometimes a drug evokes an unusual response that is *qualitatively* different from the expected response. Such an unexpected response is called a *meta-reaction*. A not uncommon meta-reaction is a central nervous stimulant rather than depressant effect of phenobarbital, especially in women. Pain and certain pathological states sometimes favor *meta*-reactivity. Responses that are different in infants or the aged than in young and middle-aged people are not *meta*-reactions if the response is usual in the age group. The term *idiosyncrasy* also denotes *meta*-reactivity, but the word bas been so abused that it is recommended that it be dropped. Although hypersensitivity may cause unusual effects, it is not included in *meta*-reactivity.

Drug Receptors and Receptor Theory

Most drugs act by combining with some key substance in the biological milieu that has an important regulatory function in the target organ or tissue. This biological partner of the drug goes by the name of *receptive substance* or *drug receptor*. The receptive substance is considered mostly to be a cellular constituent, although in a few instances it may be extracellular, as the cholinesterases are, in part. The receptive substance is thought of as having a special chemical affinity and structural requirements for the drug. Drugs such as emollients, which have a physical rather than chemical basis for their action, obviously do not act upon receptors. Drugs such as demulcents and astringents, which act in a nonselective or nonspecific chemical way, also are not considered to act upon receptors, since the candidate receptors have neither sharp chemical nor biological definition. Even antacids, which react with the extremely well-defined hydronium ion, cannot be said to have a receptor, since the reactive proton has no permanent biological residence.

Because of early preoccupation with physical theories of action and the classical and illogical dichotomy of chemical

and physical molecular interaction, there is a reluctance to admit receptors for drugs such as local anesthetics, general anesthetics, certain electrolytes, etc, which generally are not accepted to combine selectively with distinct cellular or organelle membrane constituents. The word receptor often is used inconsistently and intuitively. However, the term is a legitimate symbol for that biological structure with which a drug interacts to initiate a response. Ignorance of the identities of many receptors does not detract from, but rather increases, the importance of the term and general concept.

Once a receptor is identified, it frequently is no longer thought of as a receptor, although such identification may afford the basis of profound advances in receptor theory. Since the effects of anticholinesterases are derived only indirectly from inhibition of cholinesterase and no drugs are known that stimulate the enzyme, it may be argued that it is not a receptor. Nevertheless, a number of drugs ultimately act indirectly through the inhibition of such modulator enzymes and it is important for the theoretician to develop models based upon such indirect interrelations.

Enzymes, of course, readily suggest themselves as candidates for receptors. However, there is more to cellular function than enzymes. Receptors may be membrane or intracellular constituents that govern: the spatial orientation of enzymes, gene expression, compartmentalization of the cytoplasm, contractile or compliant properties of subcellular structures or permeability and electrical properties of membranes. For nearly every cellular constituent there can be imagined a possible way for a drug to affect its function; therefore, few cellular constituents can be dismissed *a priori* as possible receptors. All the receptors for neurotransmitters and autonomic agonists are membrane proteins with agonist-binding groups projecting into the extracellular space. The transducing apparatus, whereby an occupied receptor elicits a response, is called a *coupling system*. Excitatory neurotransmitters in the central nervous system, and nicotinic receptors elsewhere, are coupled to ion channels which, when opened, permit the rapid ingress, especially of sodium ions. GABA (γ-amino-butyric acid) and glycine are coupled to inhibitory chloride channels. Benzodiazepine receptors are coupled to the GABA-receptor. Beta-adrenergic receptors and a number of receptors for polypeptide hormones interact with a stimulatory GDP/GTP-binding protein (G-protein) which can activate the enzyme adenylate cyclase. The cyclase then produces 3′,5′-cyclic AMP (cAMP) which, in turn, activates protein kinases. Other receptors interact with inhibitory G-proteins. Some receptors couple to guanylate cyclase.

Alpha-adrenergic, some muscarinic and various other receptors couple to the membrane enzyme, phospholipase-C, which cleaves inositol phosphates from phosphoinositides. The cleavage product, 1,4,5-inositol triphosphate (IP_3), then causes an increase in intracellular calcium, whereas the product, diacylglycerol (DAG), activates kinase-C. There are a number of other less ubiquitous coupling systems. Substances such as cAMP, cGMP, IP_3 and DAG are called *second messengers*.

It has been found that there may be several different receptors for a given agonist. Differences may be shown not only in the types of coupling systems and effects but also by differential binding of agonists and antagonists, desensitization kinetics, physical and chemical properties, genes and amino acid sequences. The differentiation among receptor subtypes is called *receptor classification*. Receptor subtypes are designated by Greek or Arabic alphabetical prefixes and/or numerical subscripts. There are at least two each of beta-adrenergic, histaminergic, serotoninergic, GABAergic and benzodiazepine receptors, probably three of muscarinic and alpha-adrenergic and five of opioid receptor subtypes.

Occupation and Other Theories

Drug-receptor interactions are governed by the law of mass action, a concept initiated by Langley in 1878. However, most chemical applications of mass law are concerned with the rate at which reagents disappear or products are formed, whereas receptor theory usually concerns itself with the fraction of the receptors combined with a drug, similar to theories of adsorption. The usual concept is that only when the receptor actually is occupied by the drug is its function transformed in such a way as to elicit a response. This concept has become known as the *occupation theory*. The earliest clear statement of its assumptions and formulations is often credited to Clark in 1926, but both Langley and Hill made important contributions to the theory in the first two decades of this century.

In all receptor theories, the terms agonist, partial agonist and antagonist are employed. An *agonist* is a drug that combines with a receptor to initiate a response.

In the classical occupation theory, two attributes of the drug are required: (1) *affinity*, a measure of the equilibrium constant of the drug-receptor interaction, and (2) *intrinsic activity*, or *intrinsic efficacy* (not to be confused with efficacy as intensity of effect), a measure of the ability of the drug to induce a positive change in the function of the receptor.

A *partial agonist* is a drug that can elicit some but not a maximal effect and which antagonizes an agonist. In the occupation theory it would be a drug with a favorable affinity but a low intrinsic activity.

A *competitive antagonist* is a drug that occupies a significant proportion of the receptors and thereby preempts them from reacting maximally with an agonist. In the occupation theory the prerequisite property is affinity without intrinsic activity.

A *noncompetitive antagonist* may react with the receptor in such a way as not to prevent agonist–receptor combination but to prevent the combination from initiating a response, or it may act to inhibit some subsequent event in the chain of action–effect–action–effect that leads to the final overt response.

The mathematical formulation of the receptor theories derives directly from the law of mass action and chemical kinetics. Certain assumptions are required to simplify calculations. The key assumption is that the intensity of effect is a direct linear function of the proportion of receptors occupied. The correctness of this assumption is most improbable on the basis of theoretical considerations, but empirically it appears to be a close enough approximation to be useful. A second assumption upon which formulations are based is that the drug-receptor interaction is at equilibrium. Another common assumption is that the number of molecules of receptor is negligibly small compared to that of the drug. This assumption is undoubtedly true in most instances, and departures from this situation greatly complicate the mathematical expression of drug-receptor interactions.

The first clearly stated mathematical formulation of drug–receptor kinetics was that of Clark.[1] In his equation,

$$Kx^n = \frac{y}{100 - y} \tag{2}$$

where K is the affinity constant, x is the concentration of drug, n is the molecularity of the reaction, and y is the percent of maximum response. Clark assumed that y was a linear function of the percent of receptors occupied by the drug, so that y could also symbolize the percent of receptors occupied. When the equation is rearranged to solve for y,

$$y = \frac{100Kx^n}{1 + Kx^n} \tag{3}$$

A Cartesian plot of this equation is identical in form to that shown in Fig 35-2. When y is plotted against $\log x$ instead of x, the usual sigmoid curve is obtained. Thus, it may be seen that the dose–intensity curve derives from mass action equilibrium kinetics, which in turn derive from the statistical nature of molecular interaction. The fact that dose–intensity and dose–percent curves have the same shape shows that they involve similar statistics.

If Eq 2 is put into log form

$$\log K + n \log x = \log \frac{y}{100 - y} \qquad (4)$$

a plot of $\log y/100 - y$ against $\log x$ then will yield a straight line with a slope of n; n is theoretically the number of molecules of drug which react with each molecule of receptor. At present, there are no known examples in which more than one molecule of agonist combines with a single receptor, hence, n should be equal to 1, universally. Nevertheless, n often deviates from 1; deviations occur because of cooperative interactions among receptors (*cooperativity*), *spare receptors* (see below), amplifications in the response system (*cascades*), receptor coupling to more than one sequence (eg, to both adenylate cyclase and calcium channels) and other reasons. In these departures from $n = 1$, the slope becomes a characteristic of the mechanism of action and response system.

The probability that a molecule of drug will react with a receptor is a function of the concentration of both drug and receptor. The concentration of receptor molecules cannot be manipulated like the concentration of a drug. But, as each molecule of drug combines with a receptor, the population of free receptors is diminished accordingly. If the drug is a competitive antagonist, it will diminish the probability of an agonist–receptor combination in direct proportion to the percent of receptor molecules preempted by the antagonist. Consequently, the intensity of effect will be diminished. However, the probability of agonist–receptor interaction can be increased by increasing the concentration of agonist, and the intensity of effect can be restored by appropriately larger doses of agonist. Addition of more antagonist will again diminish the response, which can, again, be overcome or *surmounted* by more agonist.

Clark showed empirically, and by theory, that as long as the ratio of antagonist to agonist was constant, the concentration of the competitive drugs could be varied over an enormous range without changing the magnitude of the response (see Fig 35-6). Since the presence of competitive antagonist only diminishes the probability of agonist–receptor combination at a given concentration of agonist and does not alter the molecularity of the reaction, it also follows that the effect of the competitive antagonist is to shift the dose–intensity curve to the right in proportion to the amount of antagonist present; neither shape nor slope of the curve is changed (see Fig 35-7). Both Figs 35-6 and 35-7 are from Clark's original paper on competitive antagonism.[1]

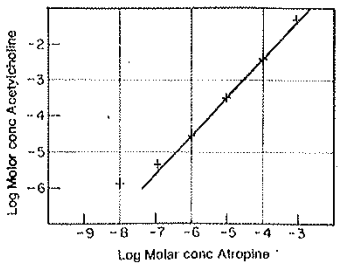


Fig 35-6. Direct proportionality of the dose of agonist (acetylcholine) to the dose of antagonist (atropine) necessary to cause a constant degree of inhibition (50%) of the response of the frog heart (courtesy, adaptation, Clark[1]).

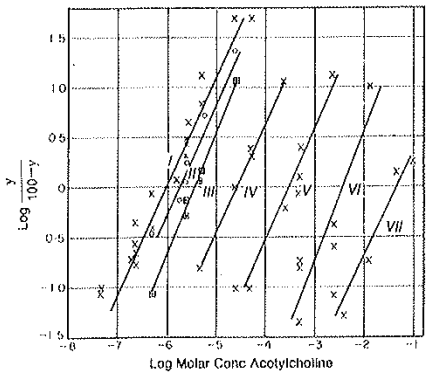


Fig 35-7. Effect of an antagonist to shift the log dose–intensity curve to the right without altering the slope. The effector is the isolated heart. *I:* no atropine; *II:* atropine, $10^{-9}M$; *III:* $10^{-7}M$; *IV:* $10^{-6}M$; *V:* $10^{-5}M$; *VI:* $10^{-4}M$; *VII:* $10^{-3}M$. *Y:* % of maximum intensity of response; the function log $y/(100\text{-}y)$ converts the log dose–intensity relationship to a straight line (courtesy, adaptation, Clark[1]).

Many refinements of the Clark formula have been made, but they will not be treated here; details and citations of relevant literature can be found among various works on receptors cited in the Bibliography. Several refinements are introduced to facilitate studies of competitive inhibition. The introduction of the concepts of intrinsic activity[2] and efficacy[3] required appropriate changes in mathematical treatment.

Another important concept has been added to the occupation theory, namely the concept of *spare receptors*. Clark assumed the maximal response to occur only when the receptors were completely occupied, which does not account for the possibility that the maximum response might be limited by some step in the action–effect sequence subsequent to receptor occupation. Work with isotopically labeled agonists and antagonists and with dose–effect kinetics has shown that the maximal effect sometimes is achieved when only a small fraction of the receptors are yet occupied. The mathematical treatment of this phenomenon has enabled theorists to explain several puzzling observations that previously appeared to contradict occupation theory.

The classical occupation theory fails to explain several phenomena satisfactorily, and it is unable to generate a realistic model of intrinsic activity and partial agonism. A rate theory, in which the intensity of response is proportional to the rate of drug–receptor interaction instead of occupation, was proposed to explain some of the phenomena that occupation theory could not, but the rate theory was unable to provide a realistic mechanistic model of response generation, and it had other serious limitations as well.

The phenomena that neither the classical occupation nor rate theory could explain can be explained by various theories in which the receptor can exist in at least two conformational states, one of which is the active one; the drug can react with one or more conformers. In a *two-state* model[4]

$$R \rightleftharpoons R^*$$

where R is the inactive and R* is the active conformer. The agonist combines mainly with R*, the partial agonist can combine with both R and R* and the antagonist can combine with R, the equilibrium being shifted according to the extent of occupation of R and R*. Other variations of occupation theory treat the receptor as an aggregate of subunits which interact cooperatively.[5]